A transposase-derived gene required for human brain development

Luz Jubierre Zapater[1,2], Elias Rodriguez-Fos[3], Merce Planas-Felix[3], Sara Lewis[4], Daniel Cameron[1,2], Phillip Demarest[1], Anika Nabila[5], Junfei Zhao[6], Paul Bergin[5], Casie Reed[1], Makiko Yamada[1,2], Alex Pagnozzi[7], Caroline Nava[8], Emilie Bourel-Ponchel[9,10], Derek E. Neilson[11], Ali Dursun[12], R. Köksal Özgül[12], Halil Tuna Akar[12], Nicholas D. Socci[1], Matthew Hayes[13], Raul Rabadan[6], David Torrents[3,14], Michael C. Kruer[3], Miklos Toth[5], Alex Kentsis[1,2,5,15]*

[1]Molecular Pharmacology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY, 10021; [2]Tow Center for Developmental Oncology, Department of Pediatrics, Memorial Sloan Kettering Cancer Center; New York, United States, 10021; [3]Barcelona Supercomputing Center (BSC), Barcelona, Spain, 08034; [4]Pediatric Movement Disorders Program, Barrow Neurological Institute, Phoenix Children's Hospital and Departments of Child Health, Neurology, Genetics and Cellular & Molecular Medicine, Phoenix, AZ; [5]Department of Pharmacology, Weill Cornell Medical College, New York, NY, 10021; [6]Program for Mathematical Genomics, Departments of Systems Biology and Biomedical Informatics, Columbia University, New York, NY; [7]CSIRO Health and Biosecurity, The Australian e-Health Research Centre, Brisbane, Australia; [8] Assistance Publique-Hôpitaux de Paris, Département de Génétique, Hôpital Pitié-Salpêtrière, Paris, France; [9]Research Group on Multimodal Analysis of Brain Function, University of Picardie Jules Verne, France; [10]Pediatric Neurophysiology Unit, Amiens Picardie University Hospital, France; [11]Phoenix Children's Hospital, Phoenix, Arizona; [12]Hacettepe University, Faculty of Medicine & Institute of Child Health, Department of Pediatric Metabolism, Ankara, Turkey; [13]Department of Physics and Computer Science, Xavier University of Louisiana, New Orleans, LA; [14]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain, [15]Departments of Pediatrics, Pharmacology, and Physiology & Biophysics, Weill Cornell Graduate School of Medical Sciences, Cornell University; New York, United States.

*Corresponding author. Email: Alex Kentsis MD, PhD; kentsisresearchgroup@gmail.com

Abstract: DNA transposable elements and transposase-derived genes are present in most living organisms, including vertebrates, but their function is largely unknown. PiggyBac Transposable Element Derived 5 (PGBD5) is an evolutionarily conserved vertebrate DNA transposase-derived gene with retained nuclease activity in cells. Vertebrate brain development is known to be associated with prominent neuronal cell death and DNA breaks, but their causes and functions are not well understood. Here, we show that PGBD5 contributes to normal brain development in mice and humans, where its deficiency causes disorder of intellectual disability, movement and seizures. In mice, Pgbd5 is required for the developmental induction of post-mitotic DNA breaks and recurrent somatic genome rearrangements in neurons. Together, these studies nominate PGBD5 as the long-hypothesized neuronal DNA nuclease required for brain function in mammals.

One-Sentence Summary: PiggyBac Transposable Element Derived 5 (PGBD5) is required for brain development in humans and mice by promoting somatic neuronal genome rearrangements.

Vertebrate brain development requires neuronal cell diversification and self-organization into signaling networks (*1*). While cell diversification is required for the development of many tissues, the development of nervous and immune systems is also uniquely dependent on DNA break repair and developmental apoptosis (*2-9*). For example, several human DNA damage repair deficiency syndromes, such as ataxia telangiectasia (AT) and Seckel syndromes, exhibit both abnormal brain neuron development and immune lymphocyte deficiencies. Likewise, mice deficient for the evolutionarily conserved DNA repair factors, such as *Xrcc5/Ku80*, are also characterized by abnormal neuron and lymphocyte development. In developing immune lymphocytes, efficient end-joining DNA repair is required to ligate DNA breaks induced by the domesticated DNA transposase RAG1/2 during somatic diversification of immunoglobulin receptor genes. Somatic genetic neuronal diversification of cell adhesion receptors was originally proposed more than 50 years ago to provide a mechanism for the complex organization of vertebrate brains (*10*) . Initially considered for clustered protocadherins based on their structural similarity to the immunoglobulin receptor genes (*11, 12*), somatic DNA breaks have now been detected in a diverse set of neuronal genes (*13, 14*). Indeed, recent single-cell sequencing studies have found extensive somatic genetic mosaicism in neurons in humans (*13, 15, 16*), as bolstered by numerous prior studies in mice (*13, 17, 18*). While somatic DNA rearrangements and developmental apoptosis are known to be essential for the evolution and function of vertebrate adaptive immunity, the mechanisms of somatic DNA breaks and neuronal apoptosis during brain development remain obscure.

Recently, we found that PiggyBac Transposable Element Derived 5 (PGBD5), the most evolutionarily conserved domesticated DNA transposase in vertebrates, is expressed in most childhood solid tumors where it mediates sequence-specific oncogenic DNA rearrangements dependent on its putative nuclease activity and cellular end-joining DNA repair (*19-21*). PGBD5-induced DNA rearrangements in human cells have been validated by multiple laboratories (*19, 22*), and recently also confirmed independently by Bigot, Yamada et al (*23*). Since most PGBD5-expressing childhood solid tumors share a common neuroectodermal developmental origin (*20, 23, 24*), we hypothesized that PGBD5 may be required for normal nervous system development by mediating somatic DNA rearrangements in developing neurons. Indeed, in both humans and mice, PGBD5 is expressed predominantly in nervous system tissues, and brain in particular (Fig. 1A-B, Fig. S1A-B).

To investigate the function of PGBD5 in human brain development, we identified two unrelated consanguineous families with *PGBD5* mutations using GeneMatcher (*25*). Exome sequencing analysis demonstrated distinct homozygous *PGBD5* mutations segregating with affected family members (Fig. 1C-D). We confirmed the observed *PGBD5* mutations using genomic PCR and Sanger sequencing (Figure S1C). *PGBD5* mutations in affected individuals involved predicted nonsense p.(Glu17*) and p.(Phe170Serfs*5) variants occurring upstream of the evolutionarily conserved aspartate triad thought to be required for the biochemical activity of the *PGBD5* DNA transposase domain in cells (Fig. 1E). The affected children were underweight with shared facial features including narrow nasal base and cupid bow upper lip. Patients presented with epilepsy and intellectual disability including impaired expressive language acquisition. All patients displayed non-progressive motor dysfunction associated with axial hypotonia, lower extremity spasticity with exaggerated deep tendon reflexes and dystonia. Brain morphometric analysis using magnetic resonance imaging (MRI) demonstrated diffuse thinning of the corpus callosum and widening of cerebellar vermis folia (Fig. 1F-G). Volumetric analysis demonstrated decreased cerebellar volume (Fig. 1F-G). While additional patients will be needed to define the full spectrum of human PGBD5 deficiency syndrome, these findings indicate that nonsense

PGBD5 variants are associated with developmental delay, intellectual disability, epilepsy, and spastic-dystonia.

To investigate the physiologic functions of Pgbd5 in nervous system development, we used dual recombinase-mediated cassette exchange to engineer *Pgbd5*$^{fl/fl}$ mice, in which *Pgbd5* exon 4 is flanked by *loxP* sites (*26*). We bred *Pgbd5*$^{fl/fl}$ mice with *EIIa-Cre* mice to generate *Pgbd5*$^{-/-}$ mice (Fig. 1H), as confirmed by genomic PCR (Fig. S1D) and Sanger sequencing. *In situ* hybridization microscopy analysis of the hippocampus, which has some of the highest density of Pgbd5-expressing neurons (Fig. 1I and Fig. S1A-B), revealed no measurable Pgbd5 transcript expression in *Pgbd5*$^{-/-}$ mice (Fig. 1I). Both *Pgbd5*$^{-/-}$ and their *Pgbd5*$^{wt/-}$ litter mate mice were born at the expected Mendelian ratios (Fig. S1E-F), but *Pgbd5*$^{-/-}$ mice were runted and had significantly smaller brains, as compared to their wild-type litter mates (t-test p = 3.4E-3 and 0.016, for females and males, respectively; Fig. 1J and Fig. S1G-J). Given the neurodevelopmental deficits associated with PGBD5 deficiency in humans, we used behavioral tests to examine *Pgbd5*$^{-/-}$ mice (Fig. 2A-B). Automated video tracking locomotor test analysis revealed significantly increased activity of *Pgbd5*-deficient mice, as compared to their wild-type litter mates (ANOVA p = 5.4E-7 and 5.5E-7, for females and males, respectively; Fig. 2A-B). Given the apparent hyperactivity of *Pgbd5*-deficient mice, we reasoned that they might have altered anxiety responses, which we assessed using the elevated plus maze test (Fig. 2C). Both female and male *Pgbd5*-deficient mice spent significantly more time and traveled longer distances in the open maze arms as compared to their wild-type litter mates, indicating reduced avoidance of the anxiogenic open arm, consistent with reduced anxiety-like behavior or the hyperactivity observed in the locomotor assay (ANOVA for phenotype p = 2.7E-6; Fig. 2D). Reduced avoidance of female and male *Pgbd5*-deficient mice was also reflected by their increased entry and time spent in the open arms (Fig. S2A-B). Prompted by the motor deficits of *PGBD5*-deficient humans, we assayed *Pgbd5*-deficient mice using the Rotarod performance test (Fig. 2E-F, Fig. S2D-E). In spite of having no significant differences in grip strength (p = 0.08; Fig. 2F), *Pgbd5*$^{-/-}$ mice exhibited significantly reduced rotarod fall latency, consistent with impaired motor learning (ANOVA p = 3.1E-7; Fig. 2E). *Pgbd5*-deficient mice also exhibited thermal hypersensitivity (Fig. S2H), without any apparent gait effects (Fig. S2F). Lastly, we assayed *Pgbd5*-deficient mice for susceptibility to seizures. We found that most *Pgbd5*$^{-/-}$ mice developed partial motor and generalized tonic-clonic seizures in response to stressful handling as compared to their wild-type litter mates (χ^2 test p = 5.8E-7; Fig. 2G). To investigate the anatomic basis of this complex behavioral syndrome, we used high-resolution manganese-enhanced MRI (MEMRI) to analyze brain architecture in *Pgbd5*-deficient mice. This revealed significant reductions in the cortical and cerebellar volumes in *Pgbd5*$^{-/-}$ mice, as assayed using quantitative volumetric mouse brain atlas analysis (t-test p = 0.033 and 8.3E-3, respectively; Fig. 2H-I, Fig. S3B-C, Fig. S3). Overall, *Pgbd5*-deficient mice display complex behavioral deficits that resemble the human PGBD5 deficiency syndrome described above, with shared seizures, behavioral and motor deficits, and structural brain abnormalities.

PiggyBac-type enzymes utilize conserved aspartate residues to catalyze DNA hydrolysis and rearrangements, with analogous residues required for the cellular DNA remodeling activities of PGBD5 (*19-21*), though whether PGBD5 functions enzymatically as a transposase, recombinase, or nuclease needs to be determined (*23*). To test whether mouse brain development requires Pgbd5 nuclease activity, we used CRISPR engineering to generate *Pgbd5*^{D262A} knock-in mice (Fig. 2J, Fig. S4A), in which one of the evolutionarily conserved aspartate residues required for cellular DNA activity was mutated to inactive alanine (*19*). We confirmed that the analogous mutation in human *PGBD5* does not impair cellular protein stability by Western immunoblotting or its ability to associate with chromatin by ChIP-seq (*19*). We verified the *Pgbd5*^{D262A} mutation

in two independent founder strains using genomic PCR and Sanger sequencing (Fig. S4A), germ-line transmission by restriction enzyme mapping (Fig. S4B), and lack of off-target gene mutations by whole-genome sequencing (Fig. 2L; Table S2). Homozygous *Pgbd5*$^{D262A;D262A}$ (*Pgbd5*$^{ki/ki}$) mice exhibited physiologic and unaltered expression of endogenous Pgbd5 mRNA as compared to their *Pgbd5*$^{wt/wt}$ litter mates, as assessed using in situ hybridization with *Pgbd5*-specific probes (Fig. 2K). Similar to *Pgbd5*$^{-/-}$ mice, *Pgbd5*$^{ki/ki}$ mice showed a trend toward smaller body weights as compared to their wild-type litter mates (Mann-Whitney test $p = 0.33$ and 0.77 for females and males, respectively; Fig. 2M-N). Importantly, *Pgbd5*$^{ki/ki}$ mice exhibited tonic-clonic seizures, similar to *Pgbd5*-deficient mice (χ^2 test $p = 7.4E-05$; Fig. 2O). Thus, brain developmental functions of Pgbd5 require specific aspartate activity of its transposase domain.

Mammalian neurogenesis occurs largely during embryonic development, with mouse cortex development peaking in the mid-to-late embryos (*27*). Thus, we analyzed cortex architecture in 14.5-day old (E14.5) embryos. Prior studies of this developmental period have also documented that post-mitotic neurons accumulate extensive DNA breaks and activate end-joining DNA repair as they migrate to the mantle layer upon differentiation of progenitor neuroblasts in the ventricular zone (*2, 3, 7*). Thus, we used a neuron-specific tubulin isoform Tuj1 as a specific marker of post-mitotic neurons (*28*), and immunofluorescence microscopy to examine post-mitotic neurons in the brain cortices of 14.5-day old embryonal (E14.5) *Pgbd5*$^{-/-}$ mice (Fig. 3A-B, Fig. S5-S9). Using γH2AX as a specific surrogate of neuronal DNA breaks (*29*), we observed a significant reduction in the number of neurons with γH2AX foci specifically among post-mitotic (Tuj1-positive), as compared to proliferating (Tuj1-negative) neuronal precursor in *Pgbd5*$^{-/-}$ mice as compared to their wild-type litter mates (t-test $p = 0.029$; Fig. 3C, Fig. S10A-B). We confirmed the specificity of this effect by analyzing the fraction of Tuj1-negative neurons in brain cortical neurons, which showed no significant differences between *Pgbd5*-deficient and wild-type litter mate brains (Fig. 3D, Fig. S14C). The observed Pgbd5-dependent neuronal γH2AX foci were specifically induced during neurogenesis in E14.5 embryos, as analysis of E12.5 brain cortices which contain only a single mantle layer revealed no significant differences (Fig. S11). Thus, Pgbd5 is specifically required for the developmental induction of DNA breaks during cortical neurogenesis.

While Pgbd5 contains an evolutionarily conserved and enzymatically active DNA nuclease domain (*19, 20*), it is possible that the observed DNA breaks in neurons occur independently of its nuclease activity. To determine whether Pgbd5-dependent neuronal DNA breaks require DNA double-strand break repair, we analyzed genetic epistasis between *Pgbd5* and *Xrcc5/Ku80*, the key factor in non-homologous end-joining (NHEJ) DNA repair (Fig. 4E). NHEJ DNA repair is required for the ligation of double-strand DNA breaks induced by many 'cut-and-paste' DNA transposase enzymes and their domesticated derivates like PGBD5 and RAG1/2 (*30, 31*). Similar to human DNA damage repair deficiency syndromes, such as AT and Seckel syndromes, *Xrcc5*$^{-/-}$ mice have neurodevelopmental defects, associated with unrepaired DNA breaks and extensive neuronal apoptosis during cortical neurogenesis, as well as severe combined immunodeficiency due to the failure to repair RAG1/2-induced DNA breaks and rearrangements in developing lymphocytes (*32*).

First, we confirmed that *Xrcc5*$^{-/-}$ mice failed to produce normal T- and B-lymphocytes, as assayed by fluorescence-activated cell scanning (FACS) using CD4/CD8 and B220/IgM-specific antibodies, respectively, as compared to their wild-type or *Pgbd5*$^{-/-}$ litter mates (Fig. S12-13). In agreement with prior studies, *Xrcc5*$^{-/-}$ mice showed a significant increase in the number of neurons with γH2AX breaks as compared to their wild-type litter mates (t-test $p = 0.028$; Fig. S14).

Remarkably, we found that *Pgbd5*$^{-/-}$*;Xrcc5*$^{-/-}$ mice had significantly reduced DNA damage in post-mitotic Tuj1-positive neurons, but not Tuj1-negative neurons which include dividing neuronal progenitors, as compared to their *Pgbd5*$^{wt/wt}$*;Xrcc5*$^{-/-}$ litter mates (t-test $p = 0.015$; Fig. 4C-D and Fig. S15-18). Commensurate with the physiologic function of developmental neuronal DNA break repair, we found that *Pgbd5*$^{-/-}$*;Xrcc5*$^{-/-}$ mice were similarly runted and failed to thrive as compared to their *Pgbd5*$^{wt/wt}$*;Xrcc5*$^{-/-}$ litter mates (Fig. S12A-H). In all, these findings indicate that Pgbd5 is required for the developmental induction of DNA breaks and normal mouse neurogenesis and brain development.

Finally, to determine whether Pgbd5 induces somatic DNA rearrangements during brain development, we used PCR-free paired-end Illumina whole-genome sequencing (WGS) of diverse anatomically dissected brain regions from multiple individual *Pgbd5*-deficient and wild-type litter mate mice. Current single-cell DNA sequencing methods enable accurate detection of single nucleotide variation, but their requirements for DNA amplification prevent the detection of larger rearrangements, such as those expected from DNA nucleases (*33*). While bulk PCR-free DNA sequencing is not sufficiently sensitive to detect DNA rearrangements occurring in single neurons, we reasoned that if Pgbd5 functions as a somatic neuronal DNA nuclease, its developmental activity would yield recurrent somatic signals in multiple diverse wild-type but not *Pgbd5*$^{-/-}$ litter mate brains via involvement of shared loci and/or sequences in bulk cell sequencing.

We tested this conjecture by analyzing somatic DNA rearrangements observed using PCR-free paired-end Illumina WGS analysis of peripheral blood mononuclear cells (PBMC) isolated from 30-day old mice (mean genome coverage 90-fold). First, we validated that our analysis was not biased by sequencing coverage (Fig. S19 A-B) and produced accurate detection of somatic DNA variants based on the allele frequencies in matched tissues (Fig. S21C-F). We analyzed resultant sequencing data using recently developed methods optimized for the accurate detection of somatic genome variation (*34-36*). Consistent with the known somatic V(D)J DNA recombination activity of RAG1/2 in blood lymphocytes, we observed somatic deletions of the *Igkj1* and *Igkj2* loci (among other immunoglobulin receptor genes) with common breakpoints in multiple sequencing reads in PBMC as compared to matched brain tissue (mean variant fraction = 0.015; Fig. S20A). Lack of apparent somatic deletions of *Igkj1* and *Igkj2* or related immunoglobulin gene loci in fetal spleens of 14.5-day old mouse embryos as compared to their matched brain tissue confirmed the specificity of this approach (Fig. S20A), consistent with the known absence of RAG1/2 activity in fetal hematopoietic cells in mouse spleen (*37*). In contrast, we observed clonal deletions of *Pgbd5* exon 4 in both adult PBMCs and fetal spleens of *Pgbd5*$^{-/-}$ mice but not in their wildtype litter mates (Fig. S20B).

Using this comparative approach to detect developmental somatic DNA rearrangements by the domesticated RAG1/2 DNA transposase in blood cells, we examined somatic genomic variation of brain tissues dissected from *Pgbd5*$^{wt/wt}$ and *Pgbd5*$^{-/-}$ litter mate mice. We performed independent analyses to quantify somatic single nucleotide variants (SNVs) and DNA rearrangements such as deletions, and then tested for their recurrence by comparing genomic locations of somatic variants in individual mice or their anatomic brain regions. We observed no significant differences in somatic SNVs in the brain tissues of both 30-day old adult and 14.5-day old embryonal *Pgbd5*$^{wt/wt}$ as compared to their *Pgbd5*$^{-/-}$ litter mate mice, consistent with their equal chronological and biological age (median allele fraction = 0.096 and 0.094 for adult *Pgbd5*$^{wt/wt}$ and *Pgbd5*$^{-/-}$ mice, respectively; Fig. S21A-B). In agreement with the stochastic nature of somatic nucleotide substitutions, most of which are due to DNA replication in proliferating tissues (*38*), we also found no genomic regions that recurrently accumulated somatic SNVs across different

brain regions or different individual mice, which have an apparently random distribution across the mouse genome (Fig. S21A-B; Table S3).

In contrast, individual adult *Pgbd5*$^{wt/wt}$ mice showed significantly more recurrent somatic DNA rearrangements both among different individual mice and their cerebella, hippocampi, and olfactory bulbs, as compared to their *Pgbd5*$^{-/-}$ litter mates. This was detected using both recurrence of somatic structural rearrangement breakpoints and their complete overlaps (24 versus 12, and 10 versus 2, respectively; χ^2 test p = 1.3E-9 and 1.6E-17, respectively; Figs. 5A-C & S21C-G). We also observed a significant increase in the total number of somatic deletions (length = 100-500 bp) in the brain tissues of *Pgbd5*$^{wt/wt}$ as compared to their *Pgbd5*$^{-/-}$ litter mates (χ^2 test p = 0.01; Fig. S22C). However, there were no significant differences in the recurrence of somatic DNA rearrangements in 14.5-day old embryonal brain tissues isolated from *Pgbd5*$^{wt/wt}$ and *Pgbd5*$^{-/-}$ litter mate embryos (Fig. 5C and Fig. S22G), consistent with the onset of Pgbd5/Xrcc5-dependent DNA break repair during this developmental period. While some of these somatic variant calls may be due to possible artifacts of Illumina sequencing, manual inspection of the sequencing reads validated 3 out 38 (8%) DNA rearrangements as somatically induced in a Pgbd5-dependent manner (Figs. 5E-G; Table S3). This value is in agreement with the permissive FDR estimate of 0.3 for this analysis (Fig. 5A-C). Finally, recurrent somatic DNA rearrangements shared among different individual mice and brain regions showed distinct genomic distribution (Figs. 5G and S21H; Table S3). While the definition of physiologic Pgbd5 genomic targets and their rearranged sequences will require the development of improved single-cell sequencing methods, we propose that the somatically rearranged genomic elements identified here represent signals of developmental physiologic Pgbd5 activity.

Although somatic genetic mosaicism has been extensively documented in diverse tissues, and somatic genetic diversification during neurogenesis was originally proposed more than 50 years ago (*39*), the existence of physiologic somatic DNA rearrangements in vertebrate brain development has not been proven so far. Here, we demonstrate that an evolutionarily conserved domesticated DNA transposase-derived gene PGBD5 is an unanticipated cause of somatic DNA rearrangements in normal neurogenesis and mammalian brain development. We provide evidence that PGBD5 is required for normal brain development in humans and mice, where its genetic inactivation constitutes the PGBD5 deficiency syndrome, characterized by developmental delay, intellectual disability, learning and motor deficits, and recurrent seizures. This function likely requires the nuclease activity of PGBD5, as evident from studies of mice engineered to express a catalytically inactive Pgbd5 nuclease mutant. We also demonstrate that *Pgbd5* is responsible for recurrent somatic DNA breaks in developing post-mitotic cortical mouse neurons, which explains the long-standing observations of the requirement of NHEJ DNA repair for mammalian brain development. Just like RAG1/2 dependent somatic genetic diversification during normal lymphocyte development, mammalian neuronal development also requires the evolutionarily conserved NHEJ DNA repair factor XRCC5/Ku80, which we now show to be associated with PGBD5-dependent neuronal genome rearrangements. While we cannot exclude the formal possibility that PGBD5 induces DNA breaks to promote cortical neuronal death occurring during the same developmental period, we provide evidence that Pgbd5-induced somatic DNA rearrangements affect specific neuronal chromosomal loci. In all, this study sets a foundation for the identification of molecular mechanisms of PGBD5 and its substrates, neuronal genetic diversification and self-organization in brain development.

PGBD5-mediated somatic neuronal DNA rearrangements may offer a genetic mechanism for neuronal group selection and developmental apoptosis, which are known to affect at least one

half of all cells produced during mammalian neurogenesis (*40-42*). Many studies have implicated DNA replication as a cause of somatic genetic brain mosaicism. However, this mechanism does not explain how DNA breaks and repair occur in post-mitotic neurons. The data presented here offer a plausible mechanism by which physiologic somatic DNA rearrangements induced by PGBD5 may contribute to somatic neuronal diversification and cellular selection as progenitor neuroblasts differentiate and exit the cell cycle and migrate from the ventricular zone to the mantle layer, where post-mitotic neurons exhibit NHEJ-dependent DNA damage and apoptosis (*2, 3, 7*). Since RAG1/2 targets distinct genomic loci in developing B- and T-lymphocytes, PGBD5 targets may also depend on neuronal differentiation and function. It is also possible that PGBD5 has additional nuclease-independent functions in nervous system development, such as those mediated by interactions with chromatin and other cellular factors.

While we favor the conclusion that PGBD5 acts directly on DNA (*19, 20, 23*), additional biochemical and structural studies will be needed to define the exact enzymatic mechanisms of PGBD5 cellular activities and their developmental regulatory factors, including the possibility that PGBD5 promotes somatic DNA rearrangements through recruitment of other nucleases and chromatin remodeling factors. Finally, we cannot exclude non-central nervous system contributions to the developmental defects observed in PGBD5-deficient mice and humans, as PGBD5 is likely expressed in peripheral nervous system and possibly other tissues such as neuroendocrine systems.

Importantly, PGBD5-dependent DNA rearrangements are not solely responsible for the physiologic requirement for DNA damage repair in nervous system function. For example, post-mitotic neurons also require XRCC1-dependent base excision/single-strand break repair due to single-strand DNA breaks induced by developmental cytosine demethylation (*43*). Recent studies have also shown that brain aging may involve additional somatic genetic processes (*44, 45*). While we used amplification-free DNA sequencing, specificity and sensitivity of bulk Illumina sequencing have intrinsic limitations, and further studies using amplification-free single-cell analyses will be needed to establish specific somatic neuronal DNA rearrangements and their functions.

Developmentally controlled DNA rearrangements have been discovered in diverse biological processes. For example, in addition to the function of the domesticated DNA transposase RAG1/2 in immunoglobulin receptor gene diversification in vertebrate lymphocytes *(46)*, the Spo11 DNA recombinase initiates recombination in eukaryotic meiosis (*47*), the Kat1 DNA transposase controls the yeast mating switching (*48*), and the PiggyMac DNA transposase mediates somatic DNA elimination during macronucleus development in ciliates (*49*). PGBD5-dependent mammalian neuronal genome rearrangements suggest that other evolutionarily conserved DNA transposases may be domesticated as developmental somatic nucleases. This would provide molecular mechanisms for genetic diversification during physiologic somatic tissue and organ development. In turn, dysregulation of these processes can cause deleterious somatic mutations, leading to disease. In the case of RAG1/2 and PGBD5, their dysregulation is associated with somatic oncogenic DNA rearrangements in blood cancers and solid tumors affecting children and young adults (*50*). Thus, dysregulation of PGBD5 functions during brain development may also contribute to the somatic DNA rearrangements in specific neurodevelopmental disorders.

References and Notes

1. G. M. Edelman, Neural Darwinism: selection and reentrant signaling in higher brain function. *Neuron* **10**, 115-125 (1993).
2. D. E. Barnes, G. Stamp, I. Rosewell, A. Denzel, T. Lindahl, Targeted disruption of the gene encoding DNA ligase IV leads to lethality in embryonic mice. *Current biology : CB* **8**, 1395-1398 (1998).
3. Y. Gao *et al.*, A targeted DNA-PKcs-null mutation reveals DNA-PK-independent functions for KU in V(D)J recombination. *Immunity* **9**, 367-376 (1998).
4. K. M. Frank *et al.*, Late embryonic lethality and impaired V(D)J recombination in mice lacking DNA ligase IV. *Nature* **396**, 173-177 (1998).
5. J. M. Sekiguchi *et al.*, Nonhomologous end-joining proteins are required for V(D)J recombination, normal growth, and neurogenesis. *Cold Spring Harb Symp Quant Biol* **64**, 169-181 (1999).
6. Y. Lee, D. E. Barnes, T. Lindahl, P. J. McKinnon, Defective neurogenesis resulting from DNA ligase IV deficiency requires Atm. *Genes Dev* **14**, 2576-2580 (2000).
7. Y. Gu *et al.*, Defective embryonic neurogenesis in Ku-deficient but not DNA-dependent protein kinase catalytic subunit-deficient mice. *Proc Natl Acad Sci U S A* **97**, 2668-2673 (2000).
8. K. M. Frank *et al.*, DNA ligase IV deficiency in mice leads to defective neurogenesis and embryonic lethality via the p53 pathway. *Mol Cell* **5**, 993-1002 (2000).
9. P. J. McKinnon, Maintaining genome stability in the nervous system. *Nat Neurosci* **16**, 1523-1529 (2013).
10. W. J. Dreyer, W. R. Gray, L. Hood, The Genetic, Molecular, and Cellular Basis of Antibody Formation: Some Facts and a Unifying Hypothesis. *Cold Spring Harbor Symposia on Quantitative Biology* **32**, 353-367 (1967).
11. Q. Wu, T. Maniatis, A striking organization of a large family of human neural cadherin-like cell adhesion genes. *Cell* **97**, 779-790 (1999).
12. X. Wang *et al.*, Gamma protocadherins are required for survival of spinal interneurons. *Neuron* **36**, 843-854 (2002).
13. P. C. Wei *et al.*, Long Neural Genes Harbor Recurrent DNA Break Clusters in Neural Stem/Progenitor Cells. *Cell* **164**, 644-655 (2016).
14. F. W. Alt, B. Schwer, DNA double-strand breaks as drivers of neural genomic change, function, and disease. *DNA Repair (Amst)* **71**, 158-163 (2018).
15. A. M. D'Gama, C. A. Walsh, Somatic mosaicism and neurodevelopmental disease. *Nat Neurosci* **21**, 1504-1514 (2018).
16. I. L. Weissman, F. H. Gage, A Mechanism for Somatic Brain Mosaicism. *Cell* **164**, 593-595 (2016).
17. C. Sun *et al.*, Mapping the Complex Genetic Landscape of Human Neurons. *bioRxiv*, (2023).
18. J. Kim *et al.*, Prevalence and mechanisms of somatic deletions in single human neurons during normal aging and in DNA repair disorders. *Nature communications* **13**, 5918 (2022).
19. A. G. Henssen *et al.*, Genomic DNA transposition induced by human PGBD5. *eLife* **4**, e10565 (2015).
20. A. G. Henssen *et al.*, PGBD5 promotes site-specific oncogenic mutations in human tumors. *Nat Genet* **49**, 1005-1014 (2017).

21. R. Keller *et al.*, Childhood cancer mutagenesis caused by a domesticated DNA transposase. *bioRxiv*, 2022.2007.2005.498128 (2022).
22. L. Helou *et al.*, The piggyBac-derived protein 5 (PGBD5) transposes both the closely and the distantly related piggyBac-like elements Tcr-pble and Ifp2. *Journal of molecular biology* **433**, 166839 (2021).
23. Y. Bigot *et al.*, Analysis of DNA transposition by DNA transposases in human cells. *bioRxiv*, 2023.2004.2026.538406 (2023).
24. A. G. Henssen *et al.*, Therapeutic targeting of PGBD5-induced DNA repair dependency in pediatric solid tumors. *Science translational medicine* **9**, (2017).
25. N. Sobreira, F. Schiettecatte, D. Valle, A. Hamosh, GeneMatcher: a matching tool for connecting investigators with an interest in the same gene. *Human mutation* **36**, 928-930 (2015).
26. M. Osterwalder *et al.*, Dual RMCE for efficient re-engineering of mouse mutant alleles. *Nat Methods* **7**, 893-895 (2010).
27. M. Götz, W. B. Huttner, The cell biology of neurogenesis. *Nature Reviews Molecular Cell Biology* **6**, 777-788 (2005).
28. J. R. Menezes, M. B. Luskin, Expression of neuron-specific tubulin defines a novel population in the proliferative layers of the developing telencephalon. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **14**, 5399-5416 (1994).
29. L. J. Mah, A. El-Osta, T. C. Karagiannis, gammaH2AX: a sensitive molecular marker of DNA damage and repair. *Leukemia* **24**, 679-686 (2010).
30. T. A. Rinehart, C. Dean, C. F. Weil, Comparative analysis of non-random DNA repair following Ac transposon excision in maize and Arabidopsis. *The Plant journal : for cell and molecular biology* **12**, 1419-1427 (1997).
31. L. E. Woodard *et al.*, Comparative analysis of the recently discovered hAT transposon TcBuster in human cells. *PLoS One* **7**, e42666 (2012).
32. A. Nussenzweig *et al.*, Requirement for Ku80 in growth and immunoglobulin V(D)J recombination. *Nature* **382**, 551-555 (1996).
33. C. Gawad, W. Koh, S. R. Quake, Single-cell genome sequencing: current state of the science. *Nat Rev Genet* **17**, 175-188 (2016).
34. T. Rausch *et al.*, DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333-i339 (2012).
35. A. McKenna *et al.*, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303 (2010).
36. K. Ye, M. H. Schulz, Q. Long, R. Apweiler, Z. Ning, Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865-2871 (2009).
37. H. Igarashi, S. C. Gregory, T. Yokota, N. Sakaguchi, P. W. Kincade, Transcription from the RAG1 locus marks the earliest lymphocyte progenitors in bone marrow. *Immunity* **17**, 117-130 (2002).
38. A. V. Nesta, D. Tafur, C. R. Beck, Hotspots of Human Mutation. *Trends Genet* **37**, 717-729 (2021).
39. W. R. Gray, W. J. Dreyer, L. Hood, Mechanism of antibody synthesis: size differences between mouse kappa chains. *Science* **155**, 465-467 (1967).
40. A. J. Blaschke, J. A. Weiner, J. Chun, Programmed cell death is a universal feature of embryonic and postnatal neuroproliferative regions throughout the central nervous system. *J Comp Neurol* **396**, 39-50 (1998).

41. S. Rakic, N. Zecevic, Programmed cell death in the developing human telencephalon. *Eur J Neurosci* **12**, 2721-2734 (2000).
42. L. Lossi, A. Merighi, In vivo cellular and molecular mechanisms of neuronal apoptosis in the mammalian CNS. *Prog Neurobiol* **69**, 287-312 (2003).
43. W. Wu *et al.*, Neuronal enhancers are hotspots for DNA single-strand break repair. *Nature* **593**, 440-444 (2021).
44. G. Pascarella *et al.*, Recombination of repeat elements generates somatic complexity in human genomes. *Cell* **185**, 3025-3040.e3026 (2022).
45. V. Billon *et al.*, Somatic retrotransposition in the developing rhesus macaque brain. *Genome Res* **32**, 1298-1314 (2022).
46. D. G. Schatz, Developing B-cell theories. *Nature* **400**, 614-615, 617 (1999).
47. S. Keeney, J. Lange, N. Mohibullah, Self-organization of meiotic recombination initiation: general principles and molecular pathways. *Annu Rev Genet* **48**, 187-214 (2014).
48. N. Rajaei, K. K. Chiruvella, F. Lin, S. U. Aström, Domesticated transposase Kat1 and its fossil imprints induce sexual differentiation in yeast. *Proc Natl Acad Sci U S A* **111**, 15491-15496 (2014).
49. J. Bischerour *et al.*, Six domesticated PiggyBac transposases together carry out programmed DNA elimination in Paramecium. *eLife* **7**, (2018).
50. A. Kentsis, Why do young people get cancer? *Pediatr Blood Cancer* **67**, e28335 (2020).

Acknowledgments: We thank Andrew Kung, Alejandro Gutierrez, Michael Kharas, Marc Mansour, Anton Henssen, Maria Gil Mir, Hao Zhu, Helen Mueller, Yves Bigot, Gabriella Casalena and all our laboratory members for helpful suggestions, and Sandeep Reddy, Qiangqiang Zang, Songhai Shi, Adria Pares-Palacin, Rodrigo Gularte Merida, Montserrat Puiggros, Jan Korbel, Tony Papenfuss, Guillaume Bourque, Patricia Goerner Potvin, Nicolas Robine, MSK Brain Tumor Center, Molecular Cytology, Integrated Genomics, Bioinformatics, Mouse Genetics, and Animal Imaging core facilities for technical assistance, and Maria Jasin for the gift of *Ku80*-knockout mice. AK is a Scholar of the Leukemia & Lymphoma Society, and acknowledges generous support of multiple funders listed below.

Funding:

National Institutes of Health grant R01CA214812 (AK)

National Institutes of Health grant P30CA008748 (AK)

St. Baldrick's Foundation (AK)

Burroughs Wellcome Fund (AK)

Rita Allen Foundation (AK)

Pershing Square Sohn Cancer Research Alliance and The G. Harold and Leila Y. Mathers Foundation (AK)

Starr Cancer Consortium (AK)

Plan Nacional, Agencia Estatal, Spain, PID2020-119797RB-I00 (DT)

National Institutes of Health grant R35CA253126 (RR, JZ)

Author contributions:

Conceptualization: AK, LJZ

Methodology: AK, LJZ, ERF, MPF, NDS

Investigation: LJZ, ERF, MPF, SL, DC, PD, AN, PB, CR, NSB, MY

Visualization: AK, LJZ

Funding acquisition: AK

Project administration: AK, LJZ

Supervision: AK

Writing – original draft: AK, LJZ

Writing – review & editing: All authors

Competing interests: Authors declare that they have no competing interests. AK is a consultant for Novartis, Rgenta, and Blueprint. R.R. is a founder and a member of the SAB of Genotwin, and a member of the SAB of Diatech Pharmacogenetics. None of these activities are related to the work described in this manuscript.

Data and materials availability: All data are openly available via Zenodo (10.5281/zenodo.7086428), with sequencing data available from the NCBI Sequence Read Archive (PRJNA876210). Genetically-engineered mouse strains are available from the Jackson Laboratory (*Pgbd5*$^{fl/fl}$ stock number 037535).

Supplementary Materials

Materials and Methods

Supplementary Text

Figs. S1 to S22

Tables S1 to S3

References (*01–61*)

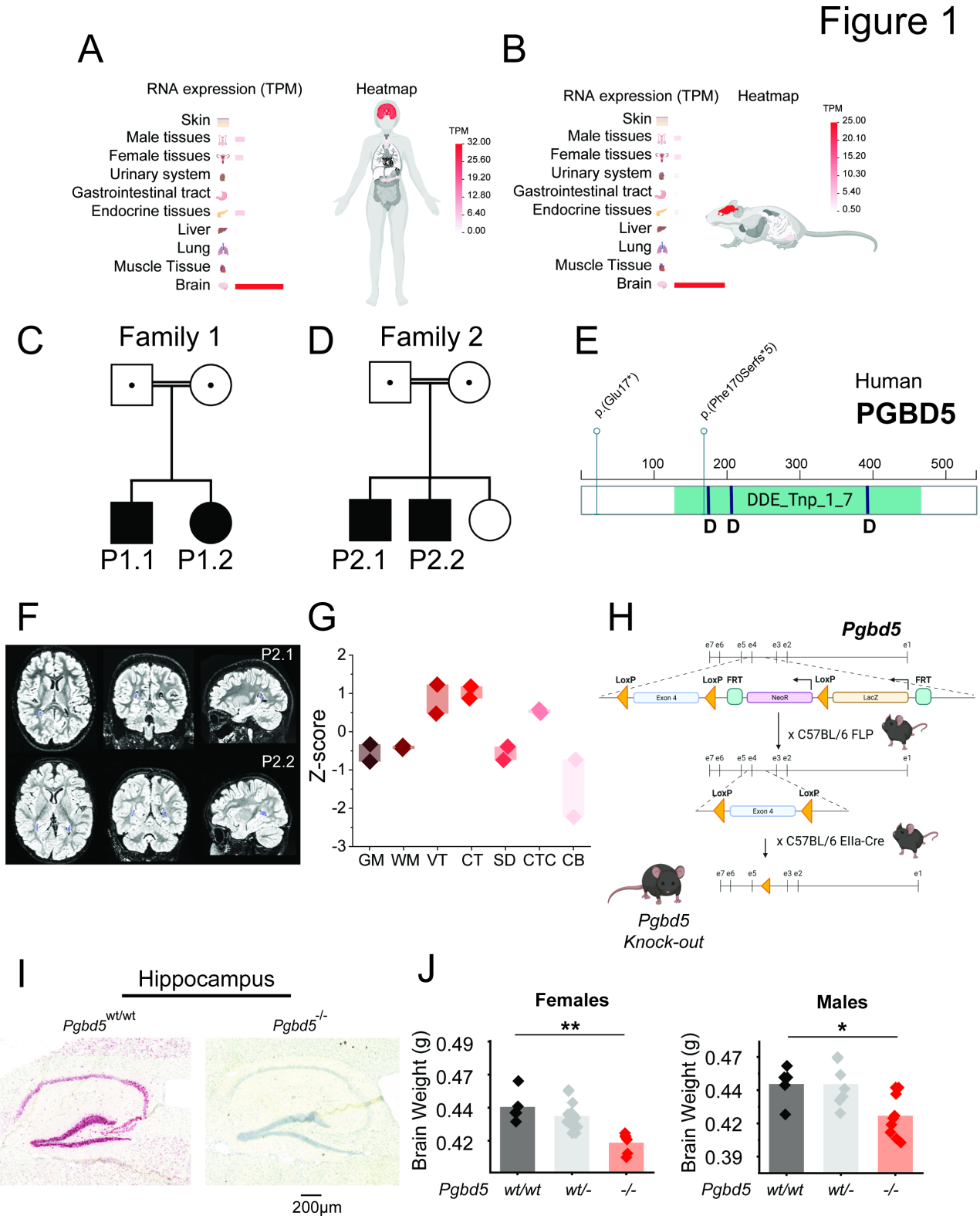

Fig. 1. *PGBD5* is specifically expressed in neuronal tissues and its deficiency causes abnormal brain development in humans. A, Bar graphs showing specific neuronal tissue expression of *Pgbd5* in **A** human and **B** mouse tissues. Color gradient from red to white indicates gene expression in transcripts per million reads (TPM). **C-D**, Segregation of *PGBD5* c.49#G>T, p.(Glu17*) and c.509del, p.(Phe170Serfs*5) alleles with disease in families 1 (**C**) and 2 (**D**), respectively. Squares

and circles indicate males and females, respectively. Dots and filled symbols indicate unaffected carriers and affected children, respectively. **E,** PGBD5 primary protein sequence schematic indicating the location of observed deleterious *PGBD5* alleles. PGBD5 transposase domain highlighted in blue, with NP_001245240.1 residues numbers indicated. DDE_Tnp_1_7 refers to the transposase homology domain. **F**, Representative T2-weighted MRI brain images from axial (left), sagittal (middle), and coronal sections (right) from family 1. Top, Patient 1-1 male imaged at age 10 y. Bottom, Patient 1-2, female imaged at 2.2 y. **G**, Box plots of *z*-scores of volumetric quantitation of specific brain regions. Cerebellum (CB) size is reduced in family 1 with a z-score representing -0.73 to -2.23 SD volume reduction compared to control population. **H**, Schematics of genetically-engineered *Pgbd5* locus and breeding strategy to generate $Pgbd5^{-/-}$ mice. Triangles and squares indicate *LoxP* and *FRT* sites, respectively, allowing for conditional excision of *NeoR* and *LacZ* targeting cassette, followed by conditional excision of exon 4 of *Pgbd5*. **I,** Representative *in situ* hybridization micrographs of the hippocampus of 60-day old mice, Pgbd5 (red) and DNA (blue), showing loss of *Pgbd5* expression in $Pgbd5^{wt/wt}$ as compared to $Pgbd5^{-/-}$ mice. Scale bar = 200 µm. **J,** Brain weights of 30-day old $Pgbd5^{wt/wt}$ (black), $Pgbd5^{wt/-}$ (grey), and $Pgbd5^{-/-}$ (red) mice, demonstrating significant reduction in brain weights in female and male mice (t-test p = 0.016 and ** p = 3.4E-3, respectively). $Pgbd5^{wt/wt}$ n=5 female and males, $Pgbd5^{wt/-}$ n=12 females and n=7 males, and $Pgbd5^{-/-}$ n=8 females and n=9 males. GM=grey matter, WM=white matter, VT=ventricle asymmetry, CT= cortical thickness, SD= sulcal depth, CTC= cortical curvature, CB= cerebellum.

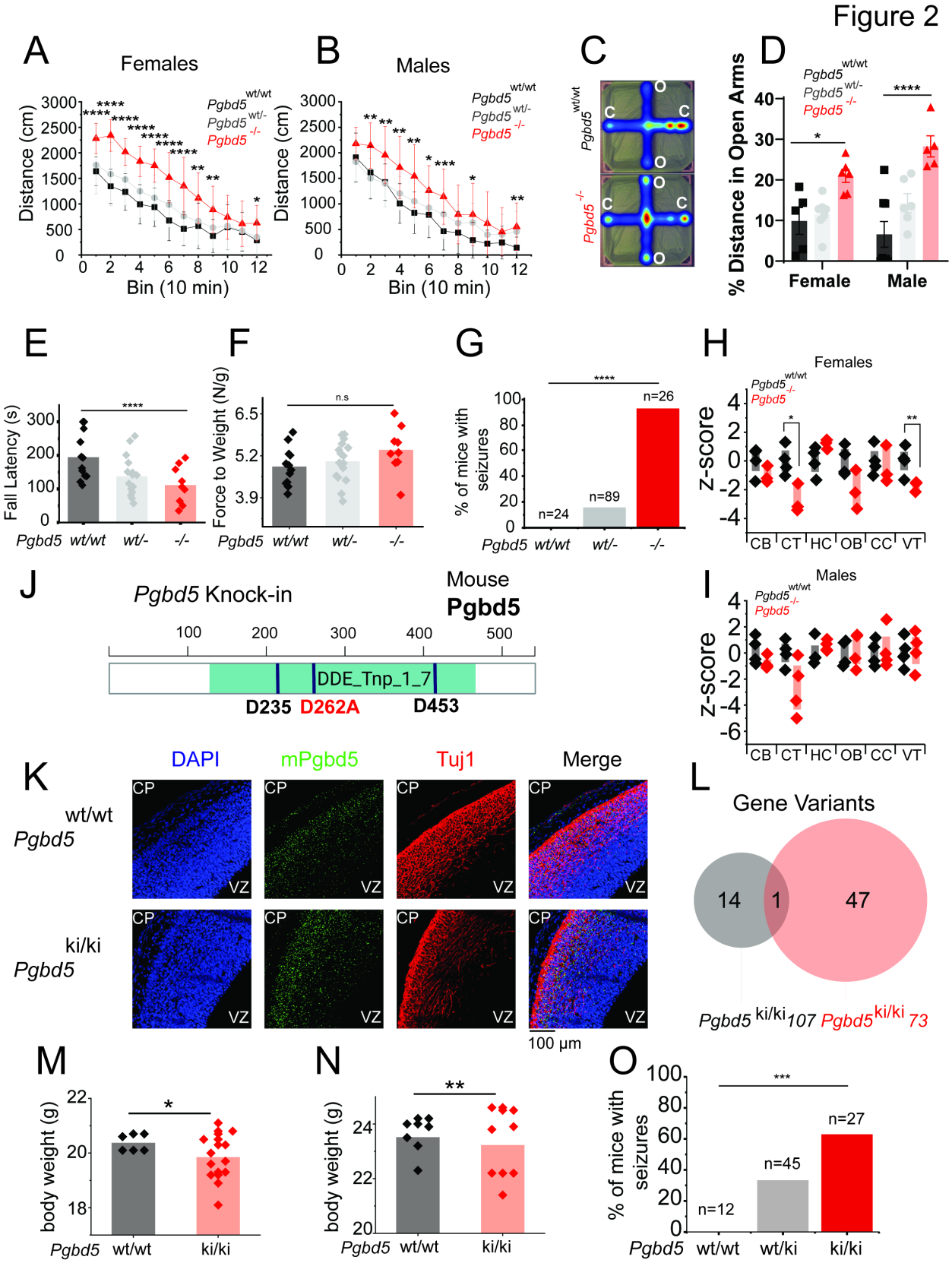

Fig. 2. *Pgbd5* knock-out mice exhibit behavioral and brain developmental deficits. A-B, Whisker plot analysis of the distance traveled in locomotor assays of 12-week old female (**A**) and

male (**B**) *Pgbd5*$^{wt/wt}$ (black), *Pgbd5*$^{wt/-}$ (grey), and *Pgbd5*$^{-/-}$ (red) mice, demonstrating significantly increased activity of *Pgbd5*-deficient mice (Female and male n=12 two-way ANOVA p = 5.41E-7 and n=12 5.47E-5, respectively, post-hoc Tukey test *p<0.05, ** p<0.01, ***p<0.001, ****p<0.0001). **C**, Representative heatmaps of elevated plus maze assay (color index from dark blue to red indicates time spent in the area), with **D**, bar plot of the percentage of the distance travelled in the open arm by 12-weeks old *Pgbd5*$^{wt/wt}$ (black), *Pgbd5*$^{wt/-}$ (grey), and *Pgbd5*$^{-/-}$ (red) mice. *Pgbd5*-deficient mice exhibit a significantly increased propensity to explore the open arms (Two-way ANOVA p=2.6E-6 for genotype and p=0.3 for sex, * Tukey test p = 0.019 and **** p = 2.67E-7). **E**, Bar plots of Rotarod fall latency, showing significant reduction by *Pgbd5*$^{-/-}$ (red) as compared to *Pgbd5*$^{wt/wt}$ mice (two-way ANOVA p = 3.1E-7). **F**, Grip strength assay showing lack of significant effects in *Pgbd5*-deficient mice (t-test p = 0.08). **G**, Bar plot of seizure activity of *Pgbd5*$^{-/-}$ versus *Pgbd5*$^{wt/wt}$ litter mate mice (χ^2-test p = 5.8E-7). n indicates number of mice assayed. **H-I**, Box plots of z-scores of brain MRI volumetric measurements of 60-day old *Pgbd5*$^{-/-}$ as compared to *Pgbd5*$^{wt/wt}$ mice, showing significant reduction in cortex and ventricle size brain regions in *Pgbd5*-deficient female (**H**) and male (**I**) mice (* t-test p = 0.016, ** p = 0.037). **J,** Pgbd5 primary protein sequence schematic indicating the location of conserved aspartate triad and in red, the exon 2 D262-A substitution. ENSMUST00000140012.8 Pgbd5 transposase domain highlighted in blue. **K**, Representative fluorescence *in situ* hybridization micrographs of coronal sections of heads of 14.5-day old *Pgbd5*$^{ki/ki}$ as compared to *Pgbd5*$^{wt/wt}$ litter mate embryos, showing comparable expression of *Pgbd5* transcript between wildtype and Pgbd5$^{ki/ki}$ litter mates. Green shows mouse *Pgbd5* RNA, red shows postmitotic neurons and blue shows nuclei DAPI staining. Scale bar = 100 µm. **L**, Venn diagram showing gene variants detected in the whole genome sequencing of *Pgbd5*$^{ki/ki}$ *107* and *Pgbd5*$^{ki/ki}$*73* founder lines. Only the *Pgbd5* gene variant was found to be shared between founder lines. **M-N,** Total body weight of 60-day old *Pgbd5*$^{wt/wt}$ (black), and *Pgbd5*$^{ki/ki}$(red) mice, demonstrating a trend in reduction of total weights in female (**M**) and male (**N**) mice (Mann Whitney test p = 0.33 and ** p = 0.77, respectively). **O**, Bar plot of seizure activity *Pgbd5*$^{-/-}$ versus *Pgbd5*$^{ki/ki}$ litter mate mice (χ^2-test p = 7.41E-05). n indicates number of mice assayed. O= open arm, C= closed arm, CB= cerebellum, CT= cortex, HC= hippocampus, OB= olfactory bulb, CC= corpus callosum, VT= ventricles.

Figure 3

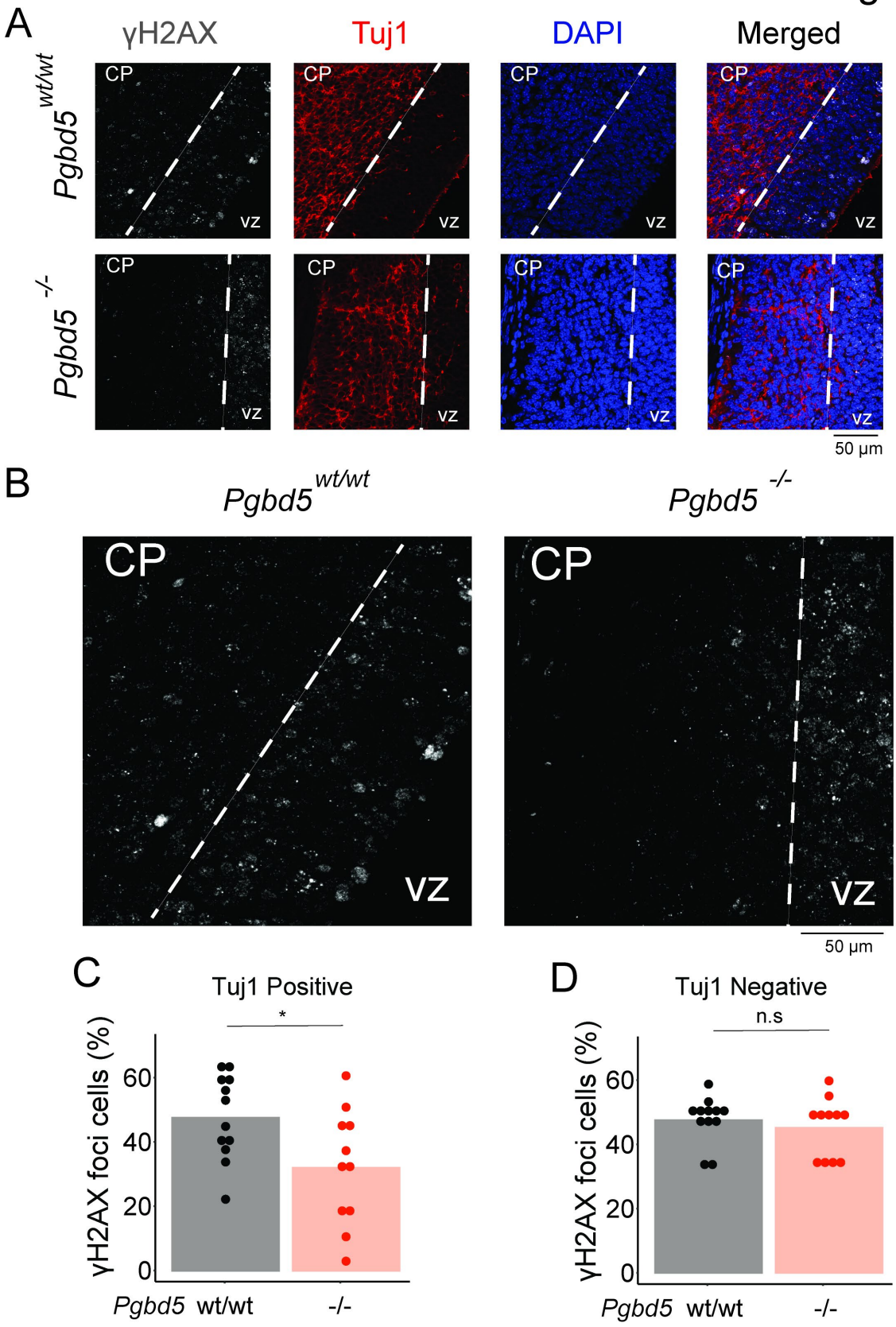
A
γH2AX
Tuj1
DAPI
Merged
Pgbd5 wt/wt
Pgbd5 -/-
CP
VZ
50 μm
B
Pgbd5 wt/wt
Pgbd5 -/-
CP
VZ
50 μm
C
Tuj1 Positive
*
γH2AX foci cells (%)
0
20
40
60
Pgbd5 wt/wt
-/-
D
Tuj1 Negative
n.s
γH2AX foci cells (%)
0
20
40
60
Pgbd5 wt/wt
-/-

Fig. 3. *Pgbd5* is required for developmental DNA breaks in postmitotic neurons. A, Representative immunofluorescence micrographs of *Pgbd5*$^{wt/wt}$ and *Pgbd5*$^{-/-}$ 14.5-day old litter mate embryos. DAPI shown in blue stains nuclei, γH2AX in white indicates sites of double strand DNA breaks, and Tuj1 in red stains differentiated postmitotic neurons. **B**, Enlarged representative immunofluorescence micrographs of *Pgbd5*$^{wt/wt}$ (left) and *Pgbd5*$^{-/-}$ (right) 14.5-day old litter mate embryos stained for γH2AX in white from panel A. **C-D**, quantification of γH2AX in postmitotic (Tuj1 positive) and proliferating neurons (Tuj1 negative). **C** and **D,** Bar plots showing percentage of cells with punctate γH2AX staining in Tuj1-positive (**C**) and Tuj1-negative neurons (**D**). In red, *Pgbd5*$^{-/-}$ mice show a significant reduction of cells with nuclear γH2AX staining (**p* = 0.029 for % of positive cells). CP= cortical plate, VZ= ventricular zone.

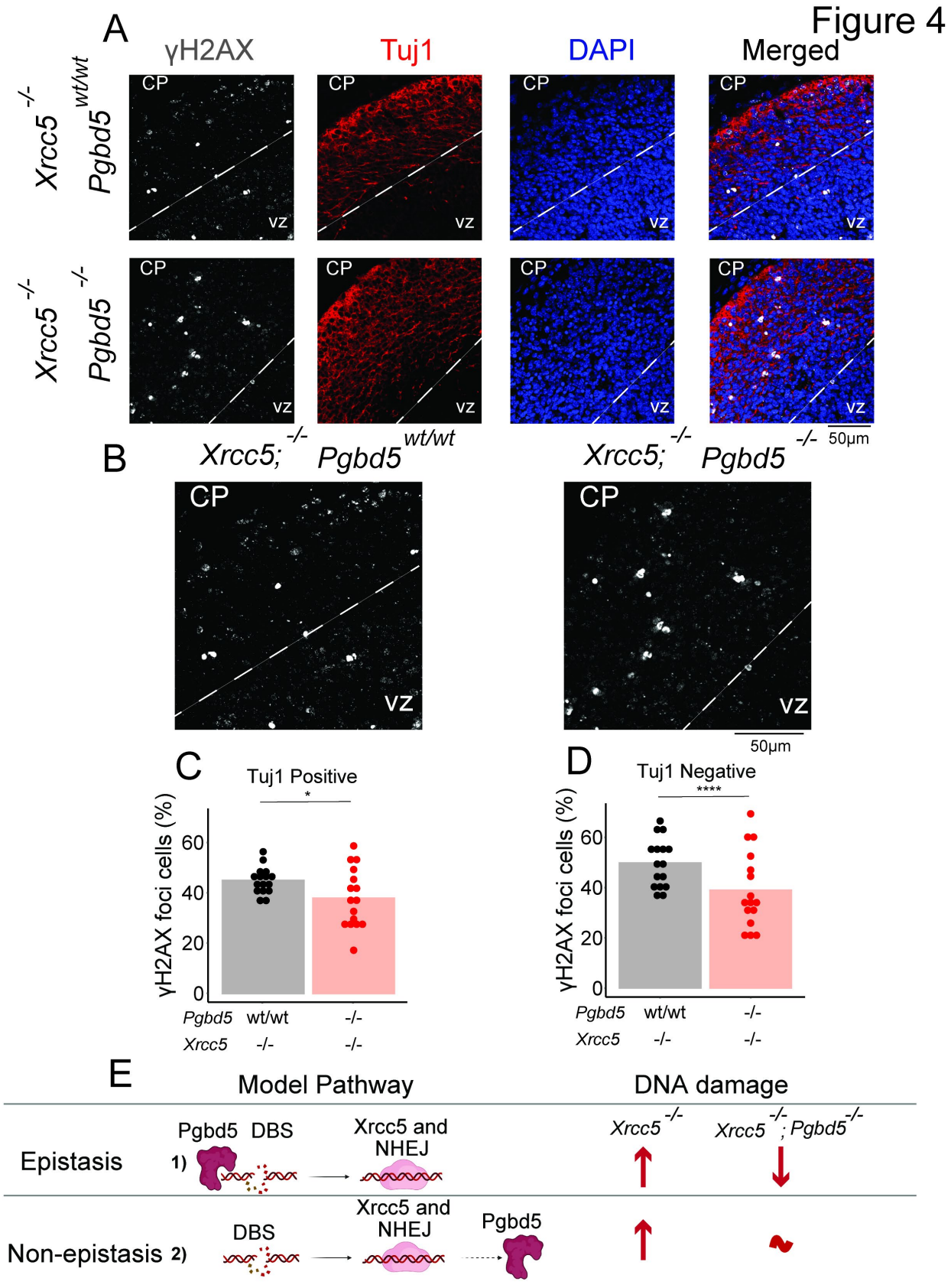
Figure 4
A
γH2AX
Tuj1
DAPI
Merged
Xrcc5 -/- Pgbd5 wt/wt
Xrcc5 -/- Pgbd5 -/-
CP
VZ
50μm
B
Xrcc5; -/- Pgbd5 wt/wt
Xrcc5; -/- Pgbd5 -/-
C
Tuj1 Positive
γH2AX foci cells (%)
Pgbd5 wt/wt -/-
Xrcc5 -/- -/-
D
Tuj1 Negative
E
Model Pathway
DNA damage
Epistasis
1)
Pgbd5 DBS
Xrcc5 and NHEJ
Xrcc5 -/-
Xrcc5 -/- ; Pgbd5 -/-
Non-epistasis 2)
DBS
Xrcc5 and NHEJ
Pgbd5

Fig. 4. *Xrcc5* is required for *Pgbd5*-induced double strand DNA break repair. **A**, Representative immunofluorescence micrographs of *Xrcc5*$^{-/-}$;*Pgbd5*$^{wt/wt}$ (up) and *Xrcc5*$^{-/-}$;*Pgbd5*$^{-/-}$ (down) 14.5-day old litter mate embryos stained for γH2AX in white. DAPI shown in blue stains nuclei, γH2AX in white indicates sites of double strand breaks, and Tuj1 in red stain postmitotic neurons. **B**, Enlarged representative immunofluorescence micrographs of *Xrcc5*$^{-/-}$;*Pgbd5*$^{wt/wt}$ and *Xrcc5*$^{-/-}$;*Pgbd5*$^{-/-}$ 14.5-day old litter mate embryos from panel A. **C-D**, Quantification of nuclear γH2AX in postmitotic neurons (Tuj1 positive) and proliferating neurons (Tuj1 negative). Bar plots showing percentages of cells with punctate γH2AX staining in Tuj1-positive (**C**) and Tuj1-negative neurons (**D;** * p = 0.015 and ****p =9.8E-5 for postmitotic and proliferating neurons, respectively). **E**, Schematic showing potential models for interactions between *Pgbd5* and *Xrcc5* in DNA break repair in developing neurons. Arrows on the right indicate DNA damage. CP= cortical plate, VZ= ventricular zone.

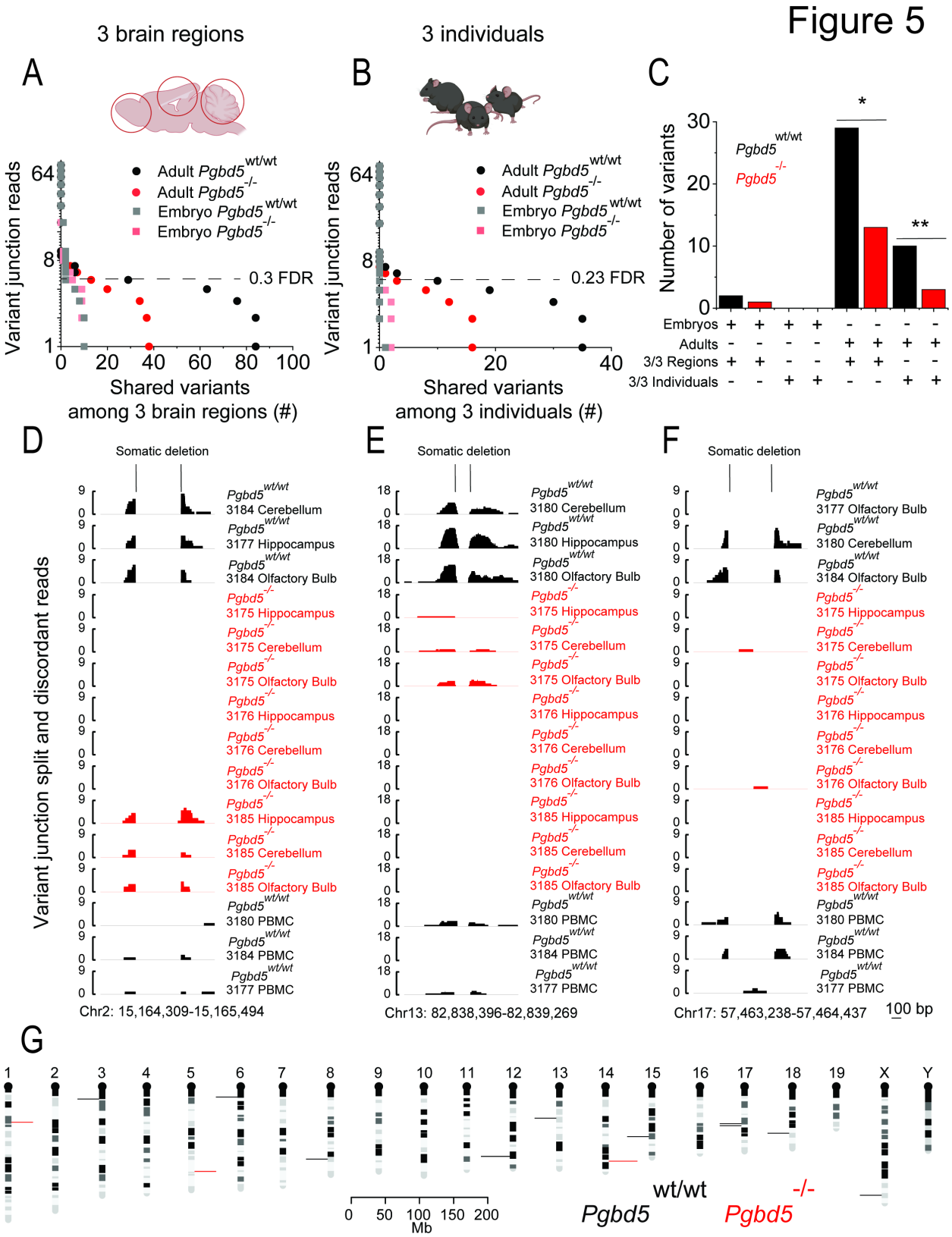

Fig. 5. *Pgbd5* is required for recurrent somatic DNA rearrangements in developing mouse brain. A-C, Dot plots showing number of structural variants at different variant junction reads thresholds shared across three brain regions (**A**) and three individuals (**B**) in adult and embryo

Pgbd5$^{wt/wt}$ (black and grey respectively) and *Pgbd5*$^{-/-}$(red and light-red respectively). The overlap among structural variants was calculated using the breakpoint analysis (Fig. S22D): 5p and 3p breakpoint ± 350bp imposing an overlap of at least 1%. There are significantly more recurrent structural variants in *Pgbd5*$^{wt/wt}$ when imposing a restrictive threshold of at least 5 variant junction reads in the recurrent events shared among 3 individuals (χ^2 test p=2.2E-145) and 3 brain regions (χ^2 test p=2.8E-133). (**C**) Bar plot summarizing the results from **A** and **B** using the threshold of 5 reads or more. Significant differences between the number of recurrent rearrangements between adult *Pgbd5*$^{wt/wt}$ and *Pgbd5*$^{-/-}$ in three individuals (χ^2 test p =1.6E-17) and three brain regions (χ^2 test p =1.3E-9) are observed. **D-F**¸ Histograms of variant junction split and discordant reads used by Delly2 to determine structural variants. Validated loci chr2:15164309-15165494 (**D**), chr13:82838396-82839269 (**E**), chr17:57463238-57464437 (**F**). In black, three recurrent *Pgbd5*$^{wt/wt}$ individuals and/or brain regions as well as the matching peripheral blood mononuclear cells (PBMC) samples. In red, histograms of *Pgbd5*$^{-/-}$ mouse brain regions. **G**, Mouse chromosomal ideogram showing the locations of recurrent somatic DNA rearrangements in three individuals observed in *Pgbd5*$^{wt/wt}$ (black) and *Pgbd5*$^{-/-}$ (red) brains binned in 1 million base intervals.

Supplementary Materials for

A transposase-derived gene required for human brain development

Luz Jubierre Zapater[1,2], Elias Rodriguez-Fos[3], Merce Planas-Felix[3], Sara Lewis[4], Daniel Cameron[1,2], Phillip Demarest[1], Anika Nabila[5], Junfei Zhao[6], Paul Bergin[5], Casie Reed[1], Makiko Yamada[1,2], Alex Pagnozzi[7], Caroline Nava[8], Emilie Bourel-Ponchel[9,10], Derek E. Neilson[11], Ali Dursun[12], R. Köksal Özgül[12], Halil Tuna Akar[12], Nicholas D. Socci[1], Matthew Hayes[13], Raul Rabadan[6], David Torrents[3,14], Michael C. Kruer[3], Miklos Toth[5], Alex Kentsis[1,2,5,15]*

[1]Molecular Pharmacology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY, 10021; [2]Tow Center for Developmental Oncology, Department of Pediatrics, Memorial Sloan Kettering Cancer Center; New York, United States, 10021; [3]Barcelona Supercomputing Center (BSC), Barcelona, Spain, 08034; [4]Pediatric Movement Disorders Program, Barrow Neurological Institute, Phoenix Children's Hospital and Departments of Child Health, Neurology, Genetics and Cellular & Molecular Medicine, Phoenix, AZ; [5]Department of Pharmacology, Weill Cornell Medical College, New York, NY, 10021; [6]Program for Mathematical Genomics, Departments of Systems Biology and Biomedical Informatics, Columbia University, New York, NY; [7]CSIRO Health and Biosecurity, The Australian e-Health Research Centre, Brisbane, Australia; [8] Assistance Publique-Hôpitaux de Paris, Département de Génétique, Hôpital Pitié-Salpêtrière, Paris, France; [9]Research Group on Multimodal Analysis of Brain Function, University of Picardie Jules Verne, France; [10]Pediatric Neurophysiology Unit, Amiens Picardie University Hospital, France; [11]Phoenix Children's Hospital, Phoenix, Arizona; [12]Hacettepe University, Faculty of Medicine & Institute of Child Health, Department of Pediatric Metabolism, Ankara, Turkey; [13]Department of Physics and Computer Science, Xavier University of Louisiana, New Orleans, LA; [14]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain, [15]Departments of Pediatrics, Pharmacology, and Physiology & Biophysics, Weill Cornell Graduate School of Medical Sciences, Cornell University; New York, United States.

Correspondence to: Alex Kentsis MD, PhD; kentsisresearchgroup@gmail.com

This PDF file includes:

Materials and Methods
Supplementary Text
Figs. S1 to S22
Tables S1 to S3

Materials and Methods

Patient recruitment, sequencing and assessment

Samples from Family 2 were analyzed using whole-exome Illumina sequencing with mean depth of coverage of 40x. A homozygous 1bp deletion resulting in a frameshift in PGBD5 was identified chr1:230492889 (hg19) NM_001258311.2 c.509 GA>G p.(Phe170SerfsTer5). An additional family was identified through GeneMatcher (*25*) with a homozygous nonsense variant identified via ES as chr1:230425880C>A (hg38) NM_001258311.1 c.49G>T p.(Glu17*). All variants had gnomAD frequency=0.Variants were additionally verified by Sanger sequencing (Fig. S1C). Family 1 Sanger sequencing performed by Integragen on genomic DNA isolated from peripheral blood using forward primer 5' GCTGGGAGGCACTGTGG and reverse primer 5' CAGCCCACGGAGAGTCTG. Family 2 Sanger sequencing performed by Functional Bio on DNA isolated from peripheral blood using Forward primer 5' TGGACAAGCTCTTACGTCCT and reverse primer 5' CCCAGGGATTGGAAATGCAG.

Clinical phenotypes were provided by collaborating physicians (CN, EBP, DA) with additional information extracted from patient photos and MRI images. Facial features assessed by D.N. Videos from Family 1 reviewed by pediatric movement disorder neurologist (MCK). Informed consent was obtained by the treating clinician under their local IRB protocol.

Gene expression analysis

Expression data from normal tissues was obtained from Genotype-Tissue Expression (GTEx) Project for humans and RIKEN FANTOM5 project for mice.

Magnetic resonance imaging

For patients, multiplanar, multisequence MRI of the brain without contrast: Sagittal T1 FLAIR, axial T2, sagittal T2 CUBE, sagittal FLAIR CUBE, axial FLAIR, axial T2*, axial DWI, coronal T1 IR 3 plane reconstructions. 1-1 acquired at ages 7 and 10 years and 1-2 acquired at age 26 months. Neuroimaging findings were also assessed by a board-certified neuroradiologist (PC) (*51*).

Several image processing steps were performed on the T1-weighted brain MRIs (*52*) including registration to the Colin 27 Average Brain Atlas, correcting image bias using the N4 algorithm, followed by intensity normalisation and image de-noising, using anisotropic diffusion. Skull stripping was performed using an in-house algorithm developed in Python. In this approach, intradural CSF was identified using thresholding and morphological operations, following which the lateral ventricles were isolated based on their spatial location, allowing the volume of the lateral ventricles (in mL) to be extracted. Cerebral brain tissues (grey matter, white matter) were then isolated based on their MR intensities using the Expectation Maximisation (EM)/Markov Random Field (MRF) approach. From the cortical grey matter segmentation, three measures of cortical shape were measured (cortical thickness, curvature and sulcal depth) to quantify shape abnormalities. Measures were converted to a z-score from healthy cortical shape measures measured from the corresponding cortical region compared to the Child Mind Institute Healthy Brain Network cohort of 564 typically developing children (TDC) (Equation 1), based on cortical regions from the Automated Anatomical Labelling (AAL) atlas.

$$z - score_{subject} = \frac{(x_{subject} - \mu_{TDC})}{\sigma_{TDC}} \quad (Eqn\ 1)$$

One MRI from each of the two patients in Family 1 passed the quality checks for initial MRI data quality and processed segmentations. For each participant, z-scores of grey matter volume, white

matter volume, ventricle asymmetry (Equation 2), was extracted.

$$(Eqn\ 2) \qquad Ventricle\ asymmetry = \frac{(vol_{left} - vol_{right})}{(vol_{left} + vol_{right})}$$

For mice, high-resolution Mn-enhanced mouse brain 3D images were acquired on a 9.4-Tesla BioSpec scanner equipped with 114-cm gradient coil (maximum gradient strength 530 mT/m; Bruker Biospin Corp., Billerica, MA). A Bruker ID 4 cm quadrature volume coil was used for RF excitation and detection. Twenty-four hours before imaging, mice were injected intraperitoneally with $MnCl_2$ at a dose of 0.5 mmol/kg. During imaging sessions mice were anesthetized with 1-2% isoflurane gas in air and were positioned prone in the scanner. Animal body temperature was maintained with a circulating warm water bath and animal respiration was monitored with an animal physiological monitoring system (SA Instruments, Inc., Stony Brook, New York). First, T2-weighted scout brain images along 3 orthogonal orientations were acquired. Then, 3D T1-weighted mouse brain images along the trans-axial orientation were acquired covering the whole brain using FLASH (Fast Low Angle Shot) gradient echo sequence with the following acquisition parameters: repetition time 22 ms, echo time 3.6 ms, and an isotropic spatial resolution of 100 μm and total imaging time of 2 hours.

Animal handling

All animal procedures were performed following the guidelines of the Institutional Animal Care and Use Committee (IACUC) and approved by Research Animal Resource Center of the Memorial Sloan Kettering Cancer Center. All mice were housed in groups with up to five animals per cage with a 12-hour light/dark cycle, starting at 06:00. Food and water were available ad libitum.

Genetic engineering of *Pgbd5*-deficient mice

To generate *Pgbd5*-deficient mice, we used the dual recombinase-mediated cassette exchange strategy (*26*). First, tm1e vector targeting exon 4 of mouse *Pgbd5* (Knockout Mouse Programme) was electroporated in C57BL/6 embryonic stem cells, followed by their microinjection into Balb/c blastocysts (Ingenious Targeting Laboratory). Resulting tm1a chimeras with a high percentage black coat color were mated to C57BL/6 FLP mice expressing the Flp recombinase to remove the Neo cassette to generate tm1c *Pgbd5*-floxed mice, as confirmed using genotyping with SC2 and SC4 primers (SC2: GAGAGCACCGTTGGTGCATATCAG, SC4: AGAGTATGAGCGGGAGAGGAGCAG). *Pgbd5*-floxed mice where crossed to B6.FVB-Tg(EIIa-cre)C5379Lmgd/J (EIIa-Cre) mice to generate tm1d *Pgbd5* deficient mice, as confirmed by genotyping with SC2 and SC4 primers. *Pgbd5*-deficient mice were backcrossed to C57BL/6J mice for six generations. *Pgbd5*-floxed mice are available from the Jackson Laboratory (strain 037535).

Genetic engineering of *Pgbd5* catalytic deficient mice

To generate *Pgbd5*-catalytic deficient mice, we used CRISPR/Cas9-assisted genome editing by zygote microinjection in C57BL/6J (The Jackson Laboratory, ME, USA). Briefly, Mice or 3-6 weeks were used as zygote donors. Fertilized eggs were recovered at pronuclear staged from oviducts of copulated females. Zygotes were microinjected in a drop of KSOM medium. Injection cocktails consisted of Cas9 protein (100 ng/μl; PNABio), in vitro synthesized Cas9 mRNA (50

ng/µl; kit, company), crRNA(s) (50 ng/µl each; Integrated DNA Technologies), tracrRNA (200 ng/µl; Integrated DNA Technologies) and donor single stranded DNA(s) (20 ng/µl each; Integrated DNA Technologies). Cocktails were mixed just before microinjection and kept on ice. Guide RNA sequences were designed by CRISPR tools (IDT, https://www.idtdna.com/site/order/designtool/index/CRISPR_SEQUENCE and ChopChop https://chopchop.cbu.uib.no/). The crRNA and tracrRNA sequences were as follows; crRNA #1 for D/A on Exon 2, sequence (AGCCACTCTGCAGGGAGTCG), crRNA #2 for D/A on Exon 3, sequence (ACATGAACCCCTGATTGACG); tracrRNA. The microinjection setup was composed of an inverted microscope (TE200-U, Nikon), microinjectors (CellTram vario, Eppendorf; Femtojet, Eppendorf) and manipulators (TransferMan 4r, Eppendorf). Injection cocktails were injected into pronuclei, and after injection, zygotes or overnight-cultured 2-cell stage embryos were surgically implanted into the oviduct of B6/CBA F1 females primed for pseudopregnancy by mating with vasectomized males. Offspring born from the implanted embryos (the founders) were screened for the presence of the insertion of the donor sequences by PCR of genomic DNA extracted from toe clips. PCR primers used are as follows; Primer #1 D/A Exon 2 (AGGCTTCTATAGCAACCGGAGCC); Primer #2 Exon 2 (TGCATGCATGGACCTGCGTGTGG); Primer #3 Exon 3(AGACTCCTGGTCAGAGAAGTCAG); Primer #4 D/A Exon 3 (TGATGAACCCGGTTGAAGAGC).

Mouse genotyping

DNA was extracted from tail clips using PureLink Genomic DNA kit (K182000, Invitrogen) following manufacturer's instructions. PCR was performed using SC2 and SC4 primers and Platinum PCR SuperMix High Fidelity (12532016, Invitrogen) reagents, following a protocol of 30 cycles that consists of denaturation at 98C for 10 sec, annealing at 55C for 30 sec and elongation at 72C for 30 sec. After the 30 cycles an additional elongation step at 72C for 10 min was performed.

Brain weight assessment

Mice were sacrifice following IACUC specifications at 60 days old. After, brain (from olfactory bulbs to medulla oblongata was extracted and weighted on a precision balance scale.

Behavioral studies

All tests were conducted using 8–12 week-old mice. During all behavioral tests, the investigators who performed the tests were blinded to experimental genotypes. Behavioral tests were conducted during the light phase, between 10:00 hours and 16:00 hours. On test days, animals were transported to the dimly illuminated behavioral laboratory and left undisturbed for one hour before testing.

Locomotor assay

The locomotor assay used a 15×21-inch black box, divided into 12 even-sized (4 × 3 inch) rectangles. The time spent and distance travelled in the two rectangles at 150 lux of light illumination were recorded and analyzed using Med Associates Inc automated tracking system, according to manufacturer's instructions (Med Associates Inc). At least 5 mice per group were used in this assay.

Elevated plus maze

The elevated plus maze used a cross maze with 12×2-inch arms with 50 lux of light illumination. Animals were introduced to the middle portion of the maze facing an open arm and

allowed to freely explore for 10 minutes. Times spent and distance travelled in the open and closed arms were measured by an automated video-tracking system (Noldus Information Technology). Data were analyzed for distance traveled in the open arm, because it is least confounded by repeated entry. At least 5 mice per group were used in this assay.

Rotarod test

An accelerating rotarod was used with 3 cm cylinders (47650, Ugo Basile). Each experiment included a training phase which included four trials with 15 min between trials. During training, animals were placed on the rod and allowed to run until the speed reached 5 rpm, then the rod was accelerated from 5 to 40 rpm over the course of 300 sec, followed by two trials for 60 seconds at 4 rpm, with 10 minutes between all trials. Mice able to remain on the rod for the training phase were assayed, following 30 minutes of rest in a cage. The final assay did not include the training phase. Only alternate rod positions were used to minimize the confounding effect of neighboring animals. At least 9 mice per group were used in this assay.

Grip strength assay

Mice were assessed for grip strength performance immediately after the rotarod test. Grip strength was measured as tension force using the force gauge (1027SM Grip Strength Meter with Single Sensor, Columbus Instruments). To assess forelimb grip strength measurement, mice were held by the base of their tails over the top of the grid so that only front paws were able to grip the grid platform T-bar. With torso in a horizontal position, mice were pulled back steadily until the grip was released, while measuring the grip force. Grip strength measurements for both sets of limbs were performed similarly, with the torso held parallel to the grid for forelimb and hindlimb measurements. Measurements were performed 5 times with 5 min resting periods. At least 9 mice per group were used in this assay.

Seizure assessment

For seizure analysis, we utilized the Racine scale, as described previously (*53*). Briefly, mice were manually suspended by their tails while rotated 10 times for 5 seconds, after which they were placed individually into empty cages for observation. Each mouse was observed for 2 minutes.

Waterbath tail immersion

Using a water bath heated at 47ºC, 1 cm of the tail tip was submerged and time to tail flick was measured. Five mice per group were used in this assay.

Tissue perfusion and fixation

Mice were anesthetized using ketamine and xylazine, followed by midline sternotomy. Upon cannulation of the left heart ventricle and connection to the Masterflex C/L peristaltic pump (Antylia Scientific), mice were perfused for two minutes with PBS, followed by six minutes with 4% (w/v) paraformaldehyde in PBS. Dissected tissues were fixed in 4% methanol-stabilized paraformaldehyde (15714-S, Electron Microscopy Sciences) in PBS for 16 hours at 4^0C. Upon anesthesia of pregnant dams and externalization of the uterus, embryos were dissected and immobilized on 6 cm plates using pin needles (Silicone Dissecting Pad Kit 501986, World Precision Instruments). Embryos were perfused with 0.5 ml of PBS, followed by 1 ml of 4% methanol-stabilized paraformaldehyde (15714-S, Electron Microscopy Sciences) in PBS. Fixed tissues were washed in deionized water twice and stored in 70% aqueous ethanol followed by paraffin embedding or incubated in a 30% (w/v) sucrose in PBS and frozen in Clear O.C.T. compound (4585, Thermo Fisher Scientific).

Immunofluorescence and Histology

Fixed specimens (at least 3 per group) where sectioned using a microtome (Leica RM2155) in 5 µm sections. Slides were deparaffinized and rehydrated using equal volumes of xylene (100%) 3 times followed by a graded alcohol series (100-50%), followed by washing with deionized water. Antigen retrieval was performed using two methods for PGBD5 antibody Proteinase K at 10U/ml or for the other antibodies citric acid-based solution (H-3300, Vector) in a 95^0C water bath for 30 min, followed by 5 min immersion in room temperature solution. Sections were then washed twice with PBS (phosphate buffer saline 0.1M, pH 7.5) and once with PBST (PBS with 0.25% Triton X-100). Slices were subsequently blocked for 30 min at room temperature (RT) using blocking solution: 10% normal donkey serum (NSD, 017-000-121, Jackson ImmunoResearch Laboratories), 1% bovine serum albumin (BSA, A2153, Sigma-Aldrich) in PBST. Primary antibodies (Table S1) were incubated for 16 hours at 4^0C in blocking solution. Slides were washed 4 times for 10 min using PBST at room temperature. Secondary antibodies (Table S1) were incubated for an hour in the dark at room temperature in 1% NSD and 1% BSA in PBST. Slides were washed 4 times in the dark using PBST for 10 min. Slides were then stained with Hoescht-33342 (H1399, Invitrogen) for 5 min. Slides were washed one last time for 5 min in PBS and coverslips were mounted at medium light using ProLong Diamond antifade mountant (P36962, ThermoFisher). Sagittal sections of cerebellum were stained with hematoxylin and eosin.

Terminal deoxynuleotidyl transferase dUTP nick end labeling (TUNEL) assay

Same deparaffination and rehydration protocol from immunofluorescence was used. Click-iT TUNEL assay was performed following manufacturer's instructions (C10617, Invitrogen).

RNA *In situ* hybridization (RNA-ISH)

Paraffin-embedded tissue sections were cut at 5 µm and kept at 4°C. Samples were loaded into Leica Bond RX, baked for 30 mins. at 60°C, dewaxed with Bond Dewax Solution (Leica, AR9222), and pretreated with EDTA-based epitope retrieval ER2 solution (Leica, AR9640) for 15 mins. at 95°C (no proteolytic retrieval). For fluorescent *in situ* hybridization (FISH), the probe for mouse *Pgbd5* (Advanced Cell Diagnostics (ACD), Cat# 561948) was hybridized for 2hrs. at 42°C. Mouse *PPIB* (ACD, Cat# 313918) and *dapB* (ACD, Cat# 312038) probes were used as positive and negative controls, respectively. The hybridized probes were detected using RNAscope 2.5 LS Reagent Kit – Brown (ACD, Cat# 322100) according to manufacturer's instructions with some modifications (DAB application was omitted and replaced with Fluorescent Alexa 488/Tyramide (Invitrogen, B40953bak) for 20 mins. at RT). FISH slides were loaded in Leica Bond RX for IF staining. Samples were pretreated with EDTA-based epitope retrieval ER2 solution (Leica, AR9640) for 20 mins. at 95°C. Mouse monoclonal antibody Tuj1 (Biolegend, MIMS-435P) was incubated for 1h at RT. Samples were then incubated with Leica Bond Post-Primary reagent (Rabbit anti-mouse linker) (included in Polymer Refine Detection Kit (Leica, DS9800)) for 8 min, followed by incubation with Leica Bond Polymer (anti-rabbit HRP) (included in Polymer Refine Detection Kit (Leica, DS9800)) for another 8 min. Alexa Fluor tyramide signal amplification reagents (CF® 594/ tyramide conjugates (Biotium 92174) were used for signal visualization. After staining, slides were washed in PBS and incubated in 5 µg/ml 4',6-diamidino-2-phenylindole (DAPI) (Sigma Aldrich) in PBS (Sigma Aldrich) for 5 min, rinsed in PBS, and mounted in Mowiol 4–88 (Calbiochem). Slides were kept overnight at -20°C before imaging.

For RNA *in situ* hybridization, BaseScope hybridization probes specific for Pgbd5 exon 4 were generated as per manufacturer's instructions (ACD, catalogue 1181898-C1). Upon cardiac

perfusion and fixation in 4% PFA/0.1M PB overnight, dissected brains were washed twice with 30% sucrose/PBS and incubated overnight at 4°C. After cryopreservation in sucrose, brains were embedded in OCT (Thermo), and blocks were sectioned sagittally in 10 micrometer sections using a rotary microtome cryostat (Leica). Sections were stored at −80°C. Cryosections were baked for one hour at 60°C and fixed in 4% PFA for 15 min followed by washing in PBS. After dehydration, epitope retrieval treatment with ER2 for 5 min at 95°C and subsequent Protease III treatment for 15 min at 40°C were performed. The probe set was hybridized for two hours at 42°C. Signal amplification steps were performed according to the manufacturer's protocol. Fast Red (Leica Bond Polymer Refine Red Detection kit DS9390) was used as chromogen. Hematoxylin was used as a counterstain. Mouse Ppib (ACD, catalogue 701078) and Bacillus subtilis dapB (ACD, catalogue 701018) probe sets were used as positive and negative controls, respectively.

Flow cytometry analysis

Thymus and spleen were collected using standard dissection. Lymphocytes were extracted by mechanical dissociation followed by erythrocyte lysis using RBC lysis buffer (Biolegend cat.420301). Cells were filtered using a 70-µm mesh (Fisherbrand cat.22-363-528). Cells were blocked on ice for 30 min using mouse IgG isotype control antibodies (Invitrogen cat. 10400C) at 1:1000 dilution in 2.5% FBS in PBS. Cells were aliquoted at $1x10^6$ cells/ml. Antibodies were added at the specified concentrations (Table S1) and incubated for 30 min on ice protected from light. Nuclear staining was done using DAPI. Flow cytometry analysis was performed using BD LSRFortessa cell analyzer and FCS Express 6 software.

Microscopy and image acquisition

Confocal imaging was performed using a Leica TCS SP5 microscope with plan-apochromat 20x/0.75NA objective lens with 4x digital zoom. Images were quantified using blinded observers. Epifluorescence and bright field images were acquired using 20x/0.8NA objective from Pannoramic P250 Flash microscope (3Dhistech).

Whole-genome sequencing

Matched litter matelitter mates (3 per group) were dissected, then, tissue was extracted, flash frozen and kept at -80ºC till further processing. For adults, olfactory bulb, hippocampus, cerebellum and peripheral mononuclear cells (PBMC) were obtained. For embryos, forebrain, midbrain, hindbrain and liver was collected. After PicoGreen quantification and quality control by Agilent BioAnalyzer, 137-500 ng of genomic DNA was sheared using a LE220-plus focused ultrasonicator (Covaris catalog # 500569) and sequencing libraries were prepared using the KAPA Hyper Prep Kit (Kapa Biosystems KK8504) kit with the following modifications to the manufacturer's instructions. After adapter ligation, libraries were subjected to 0.5X size selection using AMPure XP beads (Beckman Coulter catalog # A63882). DNA quantity was estimated by real-time PCR using sequencing adapter primers, and libraries were mixed at equimolar ratios for sequencing. Samples were sequenced using NovaSeq 6000 in a 150bp/150bp paired-end mode, with the NovaSeq 6000 SBS v1 kit and S4 flow cells (Illumina).

Sequenced data alignment and processing

All FASTQ files corresponding to the same sample were merged. Sequencing reads were aligned to the mouse reference genome (GRCm38/mm10) downloaded from (https://genome.ucsc.edu) using bwa-mem algorithm (*54*). We used bammarkduplicates to mark the duplicated reads. For the alignment summary metrics, we used Alfred v0.1.16 (*55*).

Variant calling

Variant calling was performed using Pindel version 2.2.3 (*56, 57*) and Delly2 (*34*) with the following parameters: -c 0.05, -a 0.05 and -m 15. Followed merging of results, duplicate calls were filtered using a similarity window of 300 bp, keeping those with the highest detection quality, variant allele frequency. Only variants with features exceeding default PASS quality filter were included for analysis.

Identification and analysis of somatically rearranged genes

For the detection of genes affected by observed variants we used the BEDTOOLS package (*58*) and the annotation of NCBI genes for mouse downloaded from (https://genome.ucsc.edu). To study the effect of the different subsets of variants, we used the ENSEMBL Variant Effect Predictor (https://www.ensembl.org/Tools/VEP) (*59*) with default settings.

Identification and analysis of somatic genome rearrangement regions

For the detection of genomic intervals, the regions have been created dynamically through the list of mutations contained in each group, with a static window size of 3 Mb, minimum of two mutations/window, and excluding groups of mutations that come from larger windows using Intersect in bedtools (*58*).

Recurrence analysis of polyclonal somatic DNA rearrangements

A series of R-scripts were written to first parse and filter the VCF files for events, which are available from https://github.com/kentsisresearchgroup/P5BrainReorg. The filtered events were then checked for overlaps and clusters of mutually overlapping groups were output. For the events called by the Delly2 algorithm the overlaps were computed in two ways. Method 1 looked at the actual genomic intervals and two events were defined to overlap if either the length of both was less than 100,000 bp or the percent overlap was greater than 0.1%, where the percent overlap was computed as the length of the overlap divided by the average size of the two events. Method 2 examined the overlap between the breakpoints of the somatic variants. The ends were defined as region of 701 bp centered on each end of the event. It was these edge regions that were then intersected for overlaps this time using the rule that the percent overlap was greater than 1%.

Once all overlapping events were identified, clusters were determined by looking for groups of events that all overlapped with each other. That is each element in the cluster had to overlap with every other element, as defined as a maximal clique in graph theory. We used the R package *igraph* to both construct the overlap graph and find all maximal cliques in the graph. We then filtered for sets with 3 or more events and output the table of clusters. To examine how the clusters varied on the filtering parameters this entire procedure was repeat multiple times each time adjusting one the following parameters: normalized total read counts (RC/EventSize), high-quality variant junction reads (RV), variant allele frequency (RV/[RR+RV], where RR is the high-quality reference junction reads).

For SNV and small insertions/deletions called by Mutect2, a similar procedure was used to identify overlaps for the precise same event in multiple samples. Filtering was done on the following variables: on allele depth, variant allele frequency and total depth. Since Mutect2 does not report total depth per sample we computed an estimate of total depth by taking the maximum value of the following three estimates: RD/(1-AF), AD/AF and AD+RD, where RD is the depth of the reference reads, AD is the depth of the alternate reads and AF was Mutect2's reported allele frequency. We also computed a consisted variant allele frequency of AD/(AD+RD).

To manually inspect that these rearrangements we extracted split reads and discordant pair-end reads using SAMBLASTER (*60*). We then used SparK with default -pr -cf -o -gl , and mm10 reference gene gtf file for -gtf and -gs "yes" parameters to visualize and plot the split reads found in the regions of interest (*61*).

Statistical methods

Statistical analyses were performed using OriginPro2018 or R. Statistical significance values were determined using two-tailed t-tests for continuous variables, two-tailed Fisher exact and χ^2 tests for discrete variables, and two-way ANOVA and Tukey's test for multiple comparisons analyses.

Supplemental Tables

Table S1

Antibody	*company*	*catalog number*	*concentration*
PGBD5	Abbexa	Abx019151	1:500
γH2AX	abcam	ab11174	1:500
TUJ1	BioLegend	801202	1:500
CD4	eBioscience	11-0042-82	1:500
CD8	eBioscience	25-0081-81	1:1000
B220	BioLegend	103221	1:100
IgM	eBioscience	12-5790-82	1:750
α-mouse-Alexa647	Invitrogen	A32728	1:1000
α-rabbit-Alexa555	Invitrogen	A21422	1:1000

Table S2 (separate file)

5 See file TableS2_crispr_off_targets.txt

Table S3 (separate file)

See file Sup_table_3.xlsx

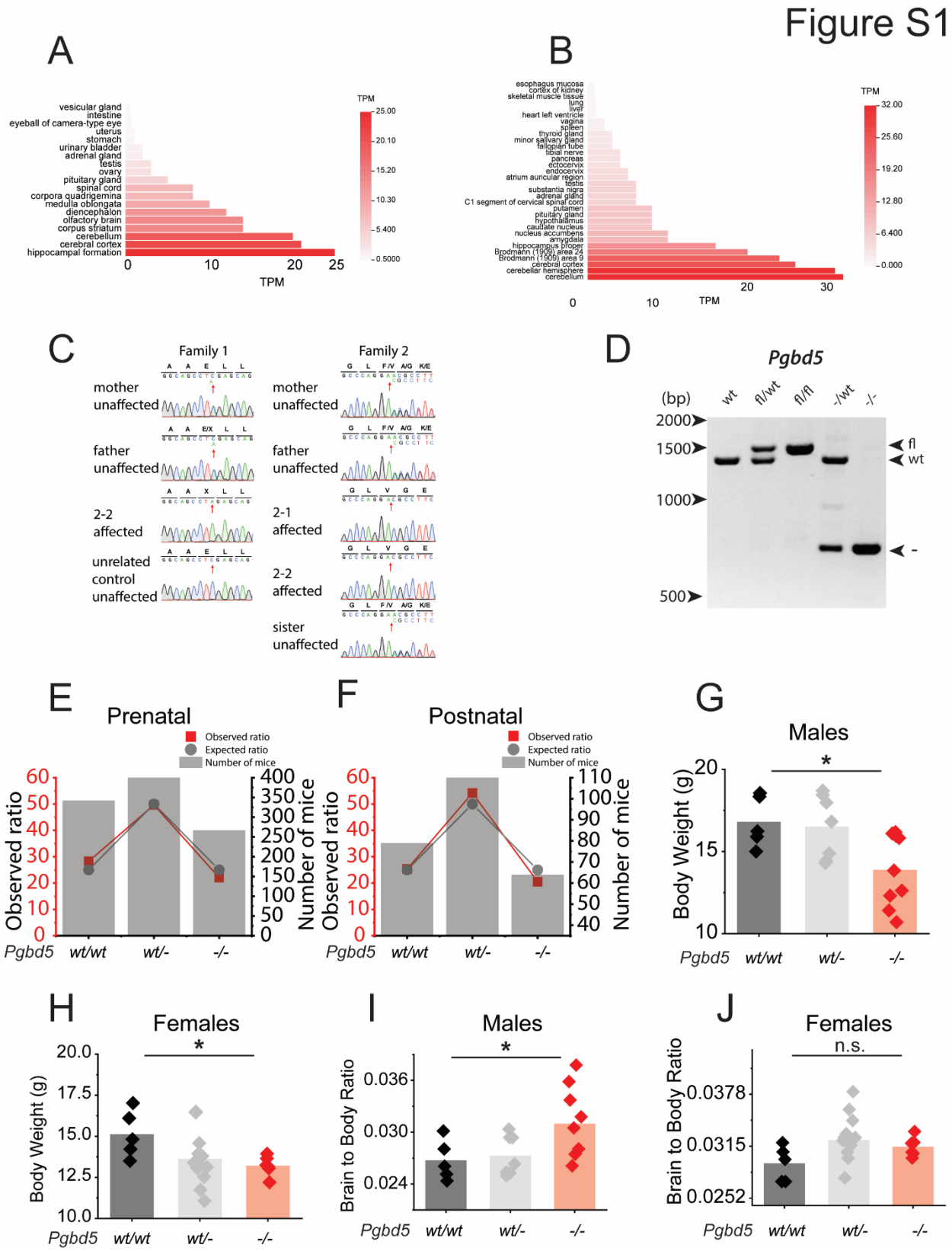

Fig. S1. *PGBD5* is highly expressed in specific regions of the brain, leading to specific phenotypic deficits when deficient in genetically-engineered *Pgbd5*-deficient mice. A-B,

Heatmap showing *PGBD5* gene expression in **A** human and **B** mouse brain regions. Highest expression is observed in the hippocampus, cerebellum and cortex. **C,** Sanger sequencing of unaffected parents and siblings, and affected children showing homozygous mutation in family 1 and 2. **D**, Gel electrophoresis analysis of genomic PCR of genetically-engineered mouse *Pgbd5* locus demonstrating floxed, wildtype and knock-out alleles. **E-F**, Mendelian ratios of *Pgbd5* $^{wt/-}$ in-crosses of prenatal (**E**) and postnatal mice (**F**). **G-H**, Body weights of female (**G**) and male (**H**) mice are significantly reduced in *Pgbd5*$^{-/-}$ mice (red); * t-test $p = 0.014$ and 0.035, respectively. **I-J**, Brain to body ratios of female (**I**) and male (**J**) mice, demonstrating significant increase in *Pgbd5*$^{-/-}$ mice (red); * t-test $p = 0.030$. **G-J** *Pgbd5*$^{wt/wt}$ n=5, *Pgbd5*$^{wt/-}$ n=8 males and n=10 females, *Pgbd5*$^{-/-}$ n=8 males and n=10 females.

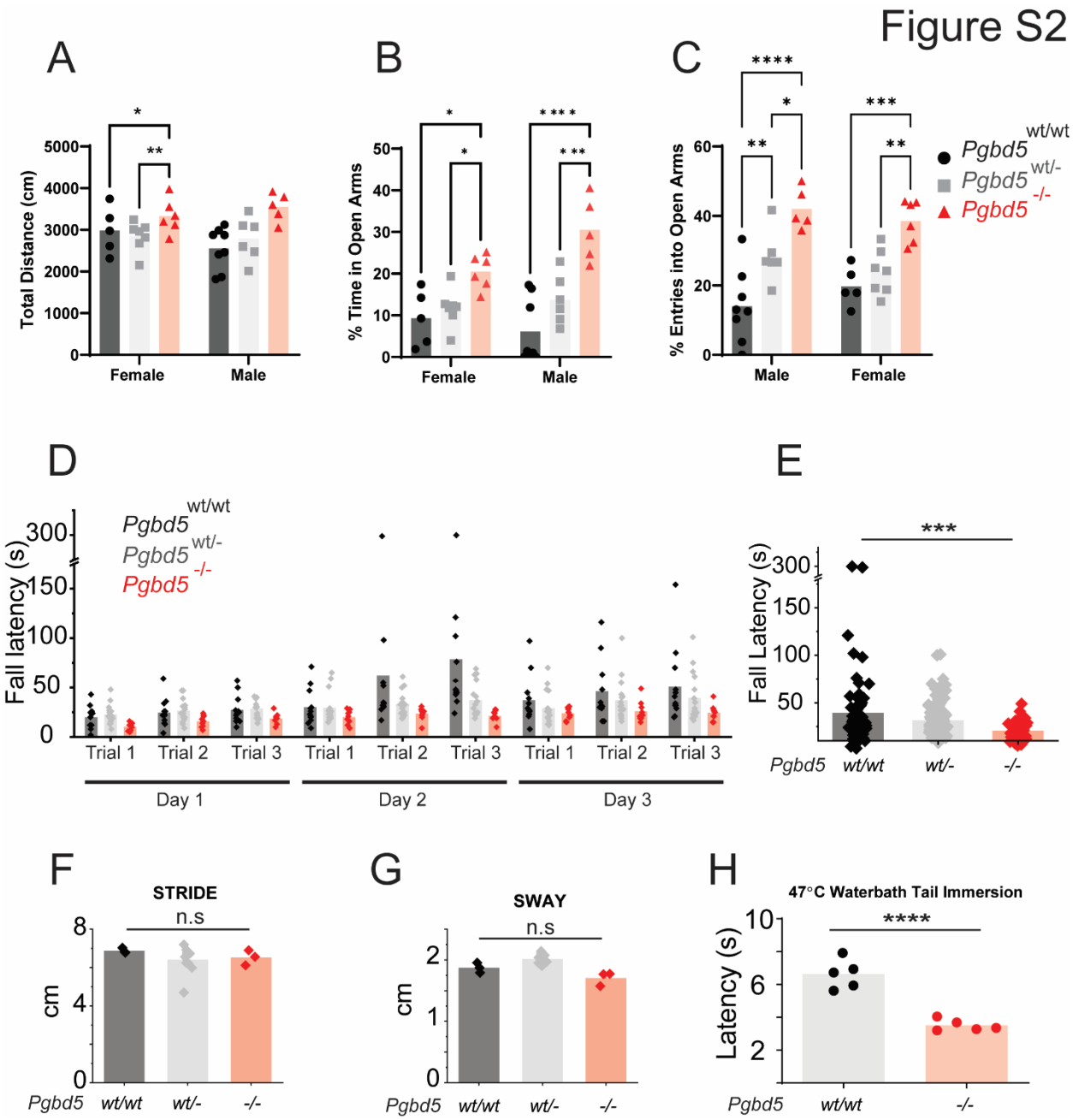

Fig. S2. *Pgbd5*-deficient mice exhibit impaired attentive behaviors and learning. A-C, Elevated-plus-maze behavioral analysis demonstrates significant increase in total distance traveled (**A**), percentage of time in open arms (**B**) and entries into open arms (**C**) of *Pgbd5*$^{-/-}$ mice (red). In total distance traveled (**A**) male *Pgbd5*$^{wt/wt}$ (black) versus *Pgbd5*$^{-/-}$ mice (red), two-way ANOVA p=1.9E-3 for phenotype and p=0.54 for sex, ** Tukey's test p = 1.6E-3; *Pgbd5*$^{-/wt}$ (grey) versus *Pgbd5*$^{wt/wt}$ (black), * Tukey's test p = 0.020. In % of time spent in open arms (**B**) male *Pgbd5*$^{wt/wt}$ (black) versus *Pgbd5*$^{-/-}$ mice (red), **** Tukey's test p = 1.0E-4; *Pgbd5*$^{-/wt}$ (grey) versus *Pgbd5*$^{-/-}$ (red), *** Tukey's test p = 3.0E-4. Females also showed significant reduction in time spent between *Pgbd5*$^{wt/wt}$ (black) and *Pgbd5*$^{-/-}$ mice (red), two-way ANOVA p= 4.1E-7 for phenotype and p= 0.17 for sex,* Tukey's test p = 0.016 and *Pgbd5*$^{-/wt}$ (grey) versus *Pgbd5*$^{-/-}$ (red), * Tukey's test p = 0.045. (**C**) All groups exhibited significant differences in time spent in open arms in male mice, *Pgbd5*$^{wt/wt}$ (black) versus *Pgbd5*$^{-/-}$ mice (red), two-way ANOVA p=9.9E-8 for phenotype

and p= 0.75 for sex, **** Tukey's test p = 1.0E-4; *Pgbd5*$^{wt/wt}$ (black) versus *Pgbd5*$^{-/wt}$ mice (grey) ** Tukey's test p = 3.8E-3; and *Pgbd5*$^{-/wt}$ (grey) versus *Pgbd5*$^{-/-}$ mice (red), * Tukey's test p = 0.013. Females also showed significant reduction between *Pgbd5*$^{wt/wt}$ (black) and *Pgbd5*$^{-/-}$ mice (red), *** Tukey's test p = 7.0E-4; *Pgbd5*$^{-/wt}$ (grey) versus *Pgbd5*$^{-/-}$ (red), ** Tukey's test p = 3.4E-3. *Pgbd5*$^{wt/wt}$ n=5, *Pgbd5*$^{wt/-}$ n=8 males and n=10 females, *Pgbd5*$^{-/-}$ n=8 males and n=10 females. **D-G**, Rotarod test analysis demonstrates impaired motor learning reflected in the relatively short fall latency from successive day trials. (**D**) Cumulative fall latency in seconds, *** t-test p = 2.17E-3 (**E**). (**F-G**) Pgbd5-deficient mice show no apparent differences in stride (**F**) and sway in cm (**G**) *Pgbd5*$^{wt/wt}$ n=9, *Pgbd5*$^{wt/-}$ n=16, *Pgbd5*$^{-/-}$ n=12. **H**, Latency in seconds from water bath tail immersion assay showing significant reduction indicating thermal hypersensitivity in *Pgbd5*$^{-/-}$ mice (red); **** t-test p = 1.0E-4.

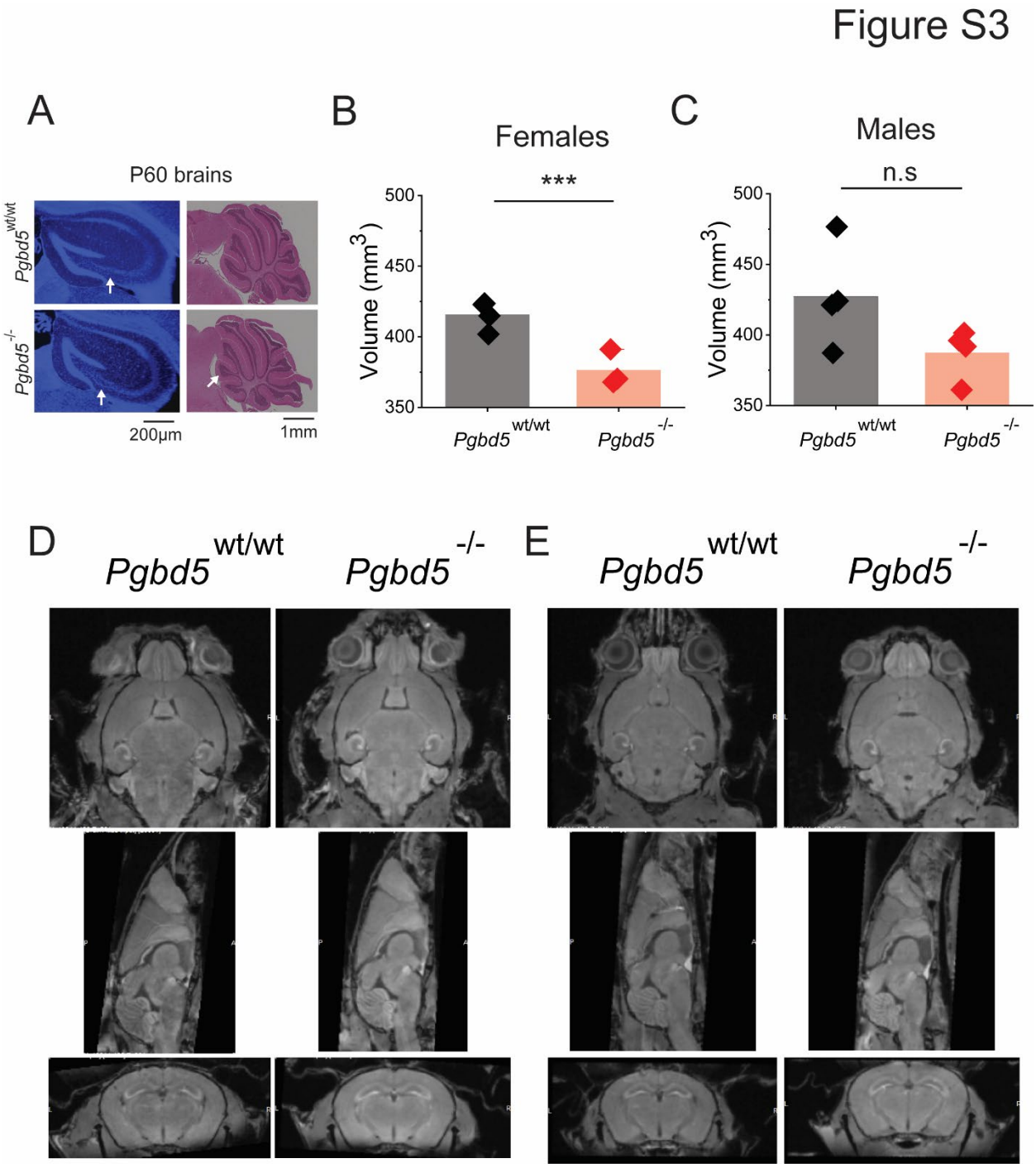

Fig. S3. *Pgbd5*-deficient mice exhibit reduced brain volumes. A, Representative fluorescence micrographs of coronal sections of P60 hippocampi *Pgbd5*$^{-/-}$ as compared to *Pgbd5*$^{wt/wt}$ litter mate mice, showing delaminated C3 (arrows). Scale bar = 200 µm and representative H&E micrographs of sagittal sections of P60 cerebellum *Pgbd5*$^{-/-}$ as compared to *Pgbd5*$^{wt/wt}$ litter mate mice, showing extra foliation in *Pgbd5*$^{-/-}$ (arrow). Scale bar = 1mm. **B-C**, Cerebral volumes in female (**B**), and male (**C**) mice as measured by microMRI; *** t-test p = 5.9E-3. **D-E**, Representative brain MRI images of male (**D**) and female (**E**) mice.

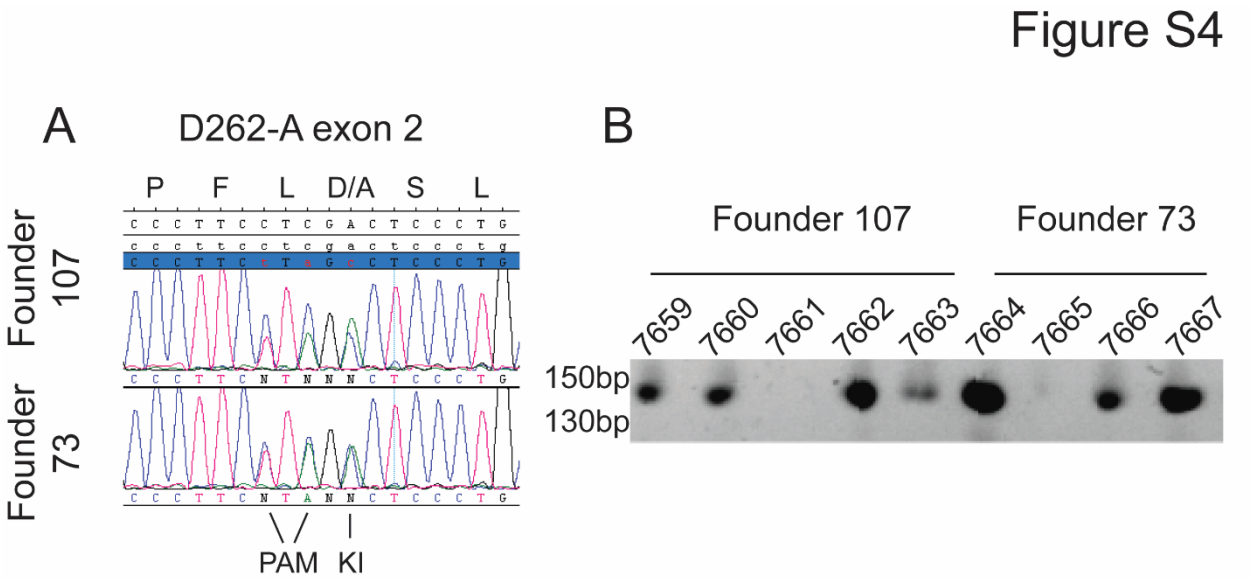

Fig. S4. Pgbd5 catalytically inactive mice phenocopy full knock-out Pgbd5 mice. A, Sanger sequence traces of the two founders showing heterozygosity of correct substitution of the PAM site and D262 codon. **B,** enzyme digestion of pcr amplicons of exon 2 D262 region shows correct transmission of both Pgbd5$^{ki/ki}$ founders upon backcrossing to C57BL/6.

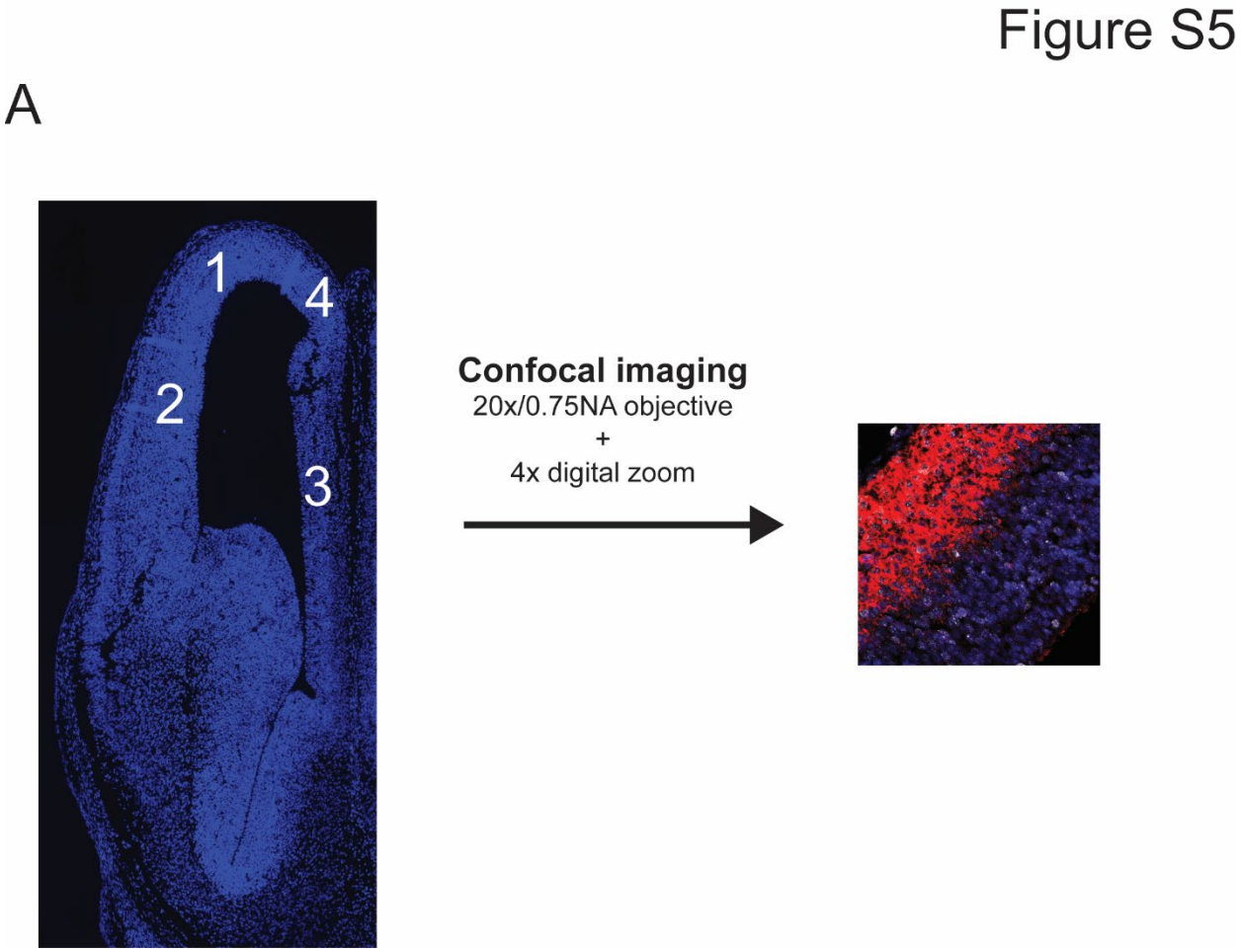

Fig. S5. Schematics on selection of cortex regions and confocal imaging acquisition. A, Representative image of a E14.5 cortex showing the 4 different cortex-regions selected and the confocal imaging process and used in Fig S6-9

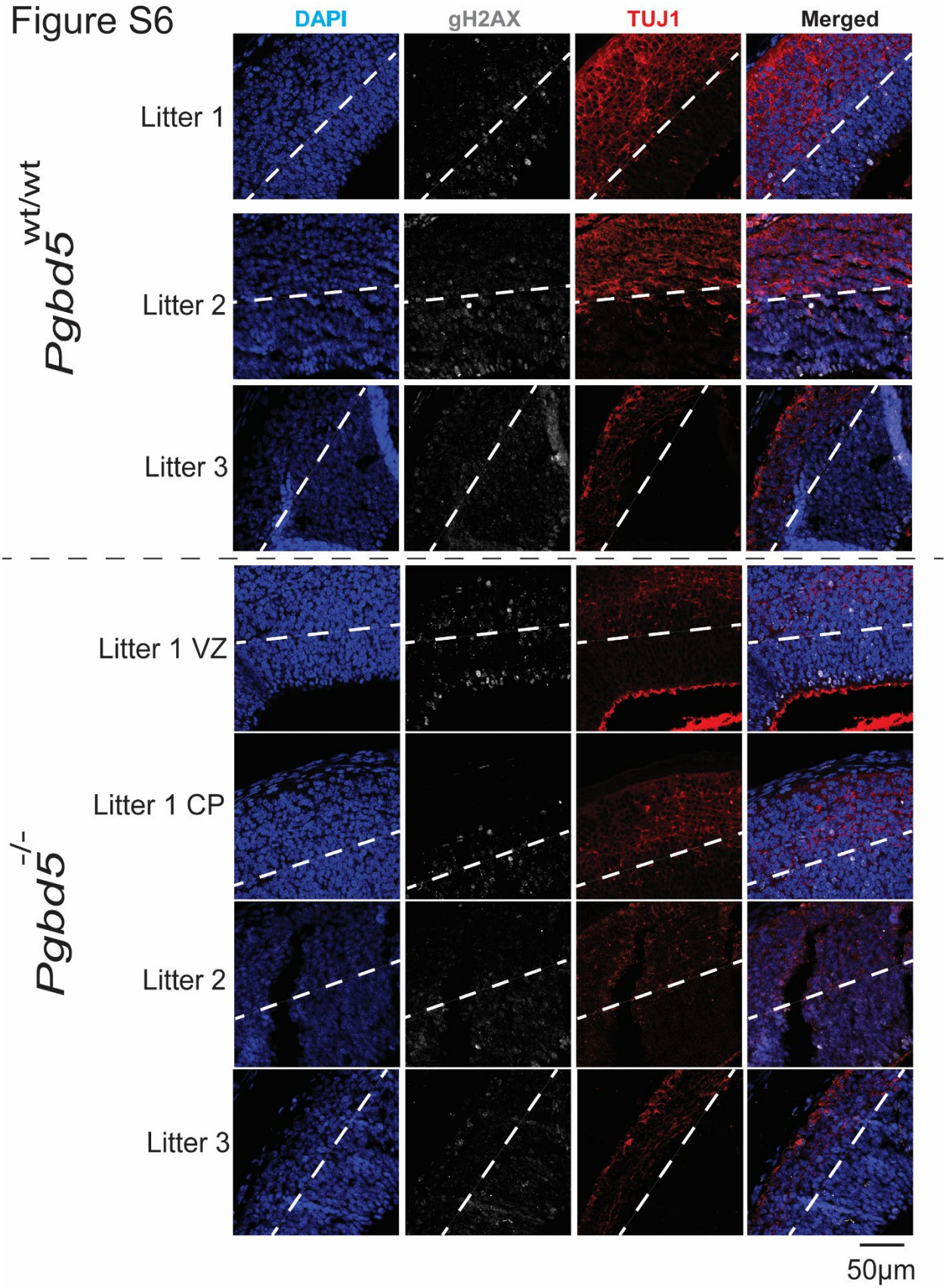

Fig. S6. *Pgbd5*-deficient mice exhibit reduced DNA damage in region of interest 1.

Immunofluorescence photographs of brain region 1 from three different independent litters, stained for DNA using DAPI (blue), gH2AX (white), and TUJ1 (red), with specific litter mates indicated (VZ, ventricular zone; CP, cortical plate). Dashed lines indicate separation between ventricular zones and cortical plates. Scale bar = 50 mm.

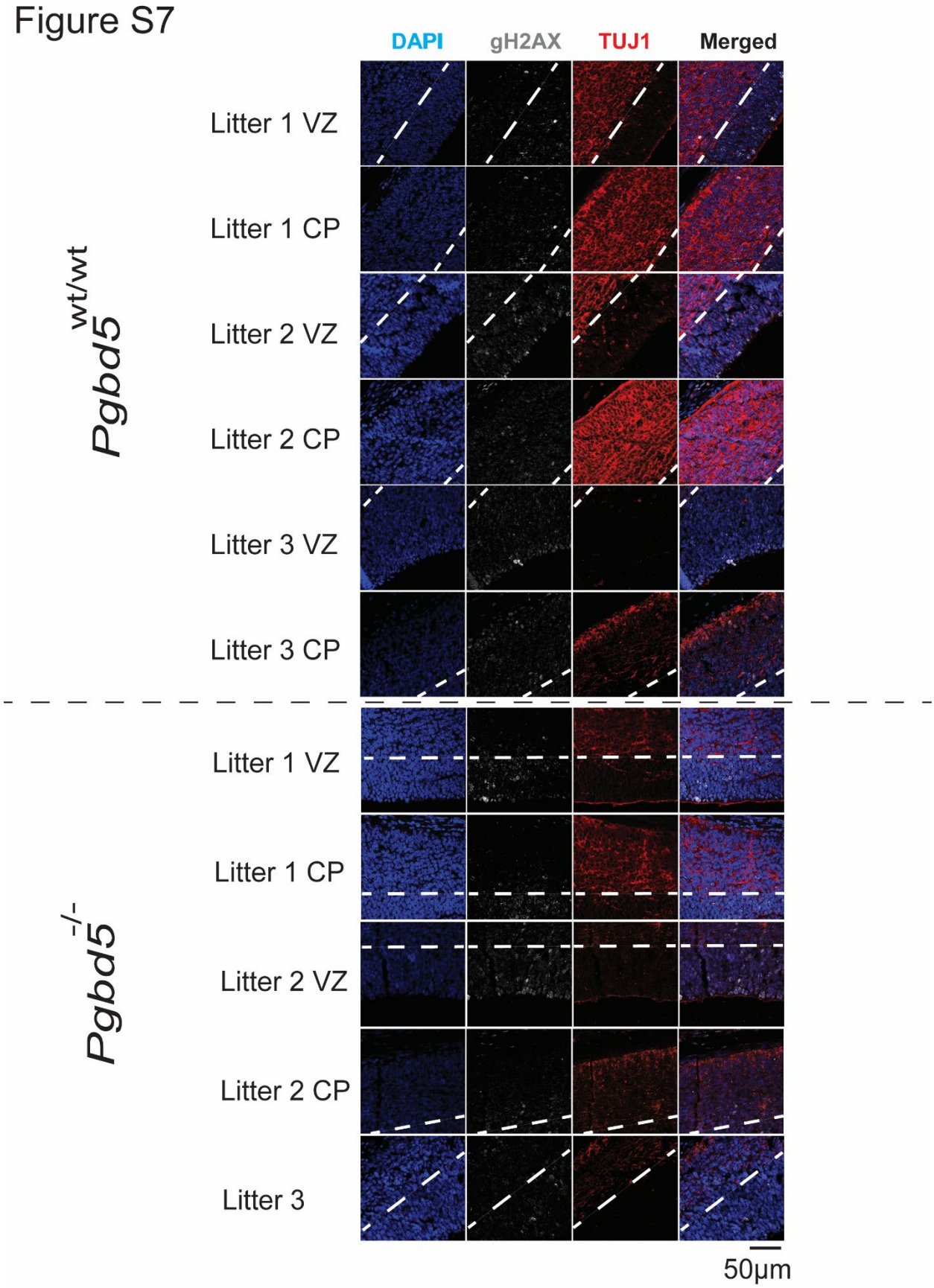

Fig. S7. *Pgbd5*-deficient mice exhibit reduced DNA damage in region of interest 2.

Immunofluorescence photographs of brain region 2 from three different independent litters, stained for DNA using DAPI (blue), gH2AX (white), and TUJ1 (red), with specific litter mates indicated (VZ, ventricular zone; CP, cortical plate). Dashed lines indicate separation between ventricular zones and cortical plates. Scale bar = 50 mm.

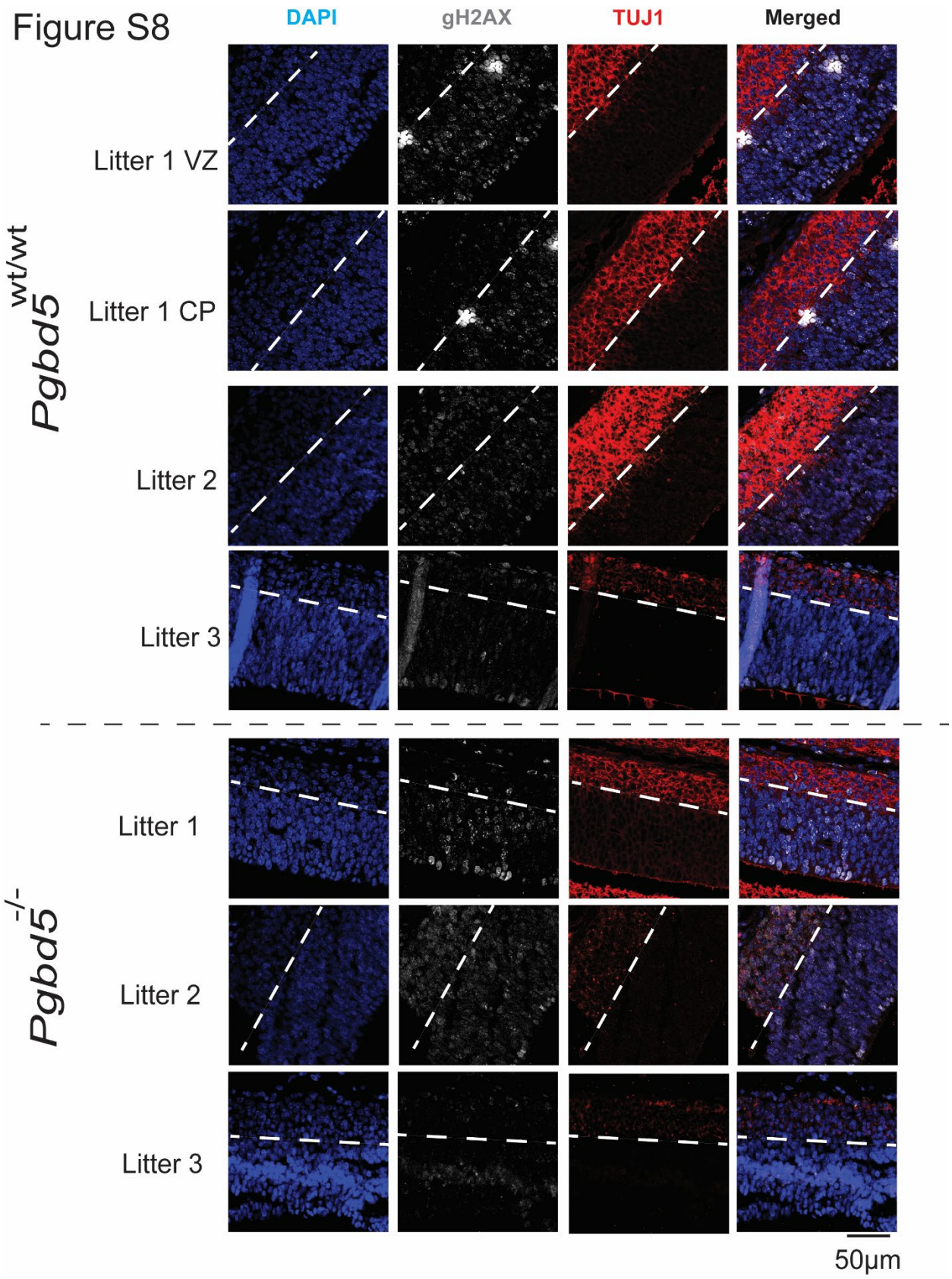

Fig. S8. *Pgbd5*-deficient mice exhibit reduced DNA damage in region of interest 3. Immunofluorescence photographs of brain region 3 from three different independent litters,

stained for DNA using DAPI (blue), gH2AX (white), and TUJ1 (red), with specific litter mates indicated (VZ, ventricular zone; CP, cortical plate). Dashed lines indicate separation between ventricular zones and cortical plates. Scale bar = 50 mm.

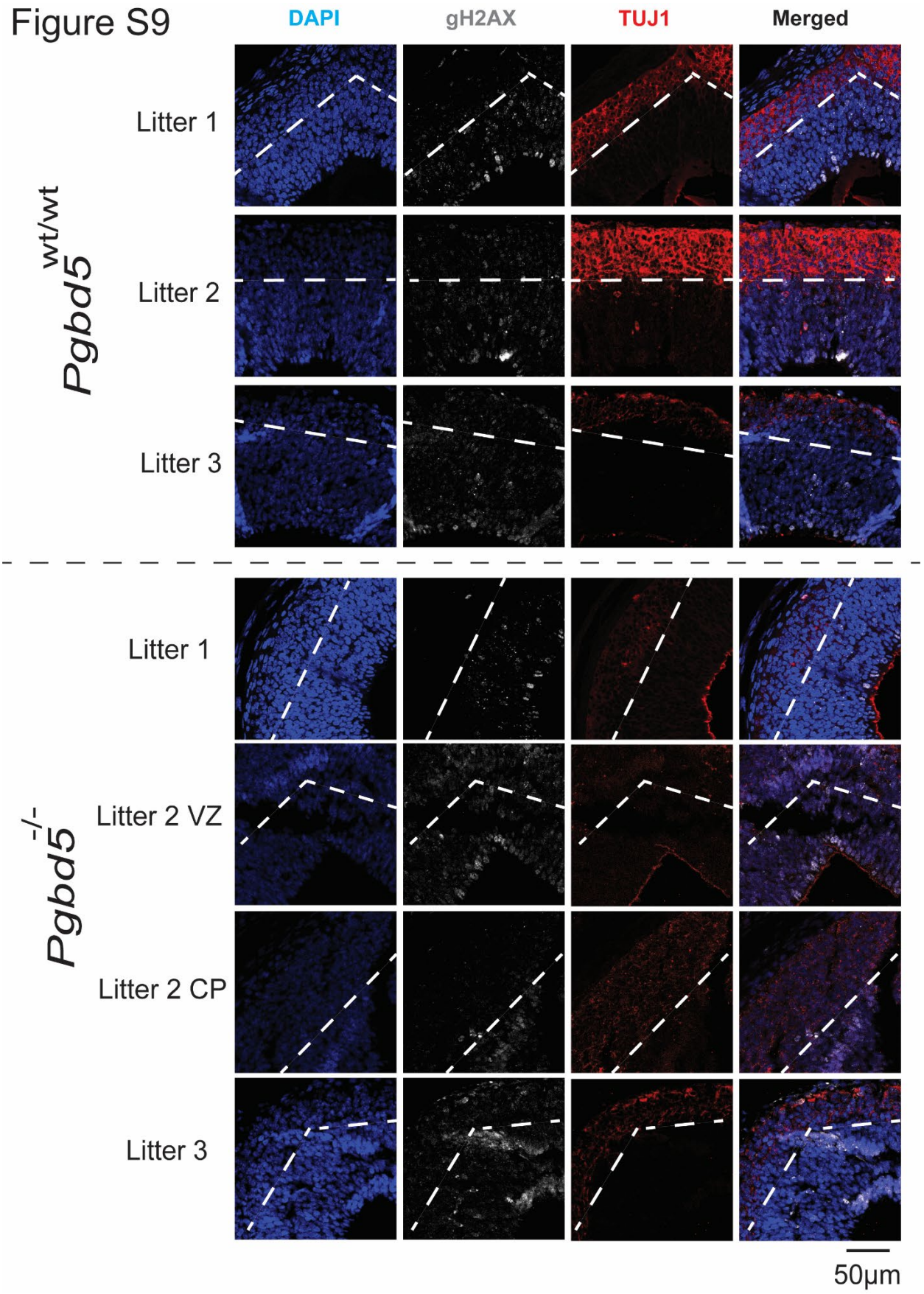
Figure S9
DAPI
gH2AX
TUJ1
Merged
Pgbd5 wt/wt
Litter 1
Litter 2
Litter 3
Pgbd5 -/-
Litter 1
Litter 2 VZ
Litter 2 CP
Litter 3
50µm

Fig. S9. *Pgbd5*-deficient mice exhibit reduced DNA damage in region of interest 4. Immunofluorescence photographs of brain region 4 from three different independent litters, stained for DNA using DAPI (blue), gH2AX (white), and TUJ1 (red), with specific litter mates indicated (VZ, ventricular zone; CP, cortical plate). Dashed lines indicate separation between ventricular zones and cortical plates. Scale bar = 50 mm.

Figure S10

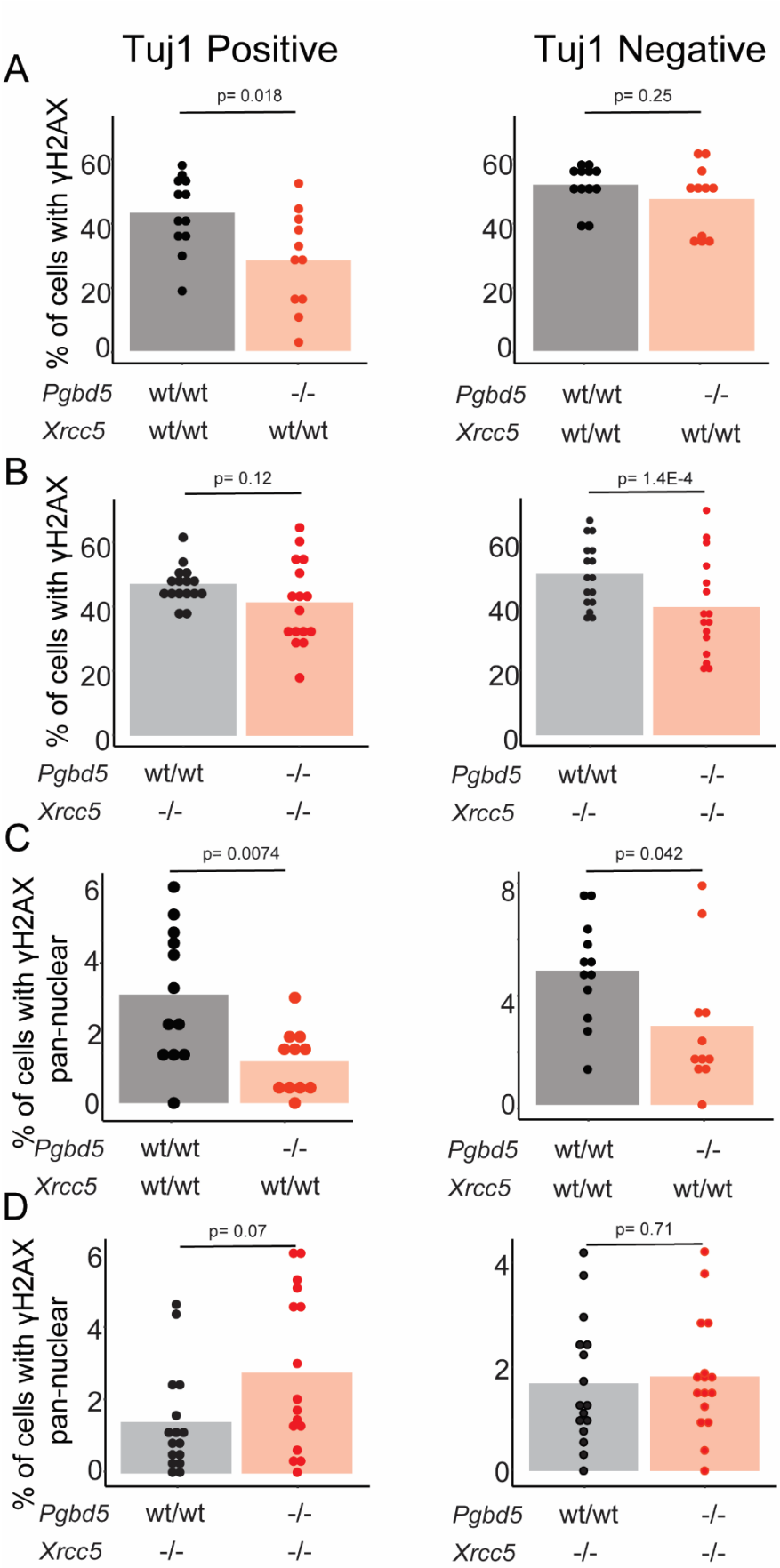
A
Tuj1 Positive
Tuj1 Negative
% of cells with γH2AX
p= 0.018
p= 0.25
Pgbd5 wt/wt -/-
Xrcc5 wt/wt wt/wt
B
p= 0.12
p= 1.4E-4
Pgbd5 wt/wt -/-
Xrcc5 -/- -/-
C
% of cells with γH2AX pan-nuclear
p= 0.0074
p= 0.042
Pgbd5 wt/wt -/-
Xrcc5 wt/wt wt/wt
D
p= 0.07
p= 0.71
Pgbd5 wt/wt -/-
Xrcc5 -/- -/-

Fig. S10. Double-mutant *Pgbd5$^{-/-}$;Xrcc5$^{-/-}$* mice exhibit reduced DNA damage as compared to single-mutant *Pgbd5$^{wt/wt}$;Xrcc5$^{-/-}$* mice, indicating epistasis of Pgbd5 expression with NHEJ DNA damage repair in E14.5 embryos. A-B, Bar plots of analysis of fraction of nuclei with punctate or pan-nuclear γH2AX in postmitotic (TUJ1 positive) and proliferative (TUJ1 negative) neurons in *Xrcc5$^{wt/wt}$;Pgbd5$^{wt/wt}$* and *Xrcc5$^{wt/wt}$;Pgbd5$^{-/-}$* (**A**) and *Xrcc5$^{-/-}$;Pgbd5$^{wt/wt}$* and *Xrcc5$^{-/-}$;Pgbd5$^{-/-}$* E14.5 mouse embryos (**B**), demonstrating a significant reduction of fraction of γH2AX-positive neurons in TUJ1-expressing cells (t-test p = 0.018). **C-D**, Bar plots of analysis of pan-nuclear γH2AX in postmitotic (TUJ1 positive) and proliferative (TUJ1 negative) neurons in *Xrcc5$^{wt/wt}$;Pgbd5$^{wt/wt}$* and *Xrcc5$^{wt/wt}$;Pgbd5$^{-/-}$* (**C**) and *Xrcc5$^{-/-}$;Pgbd5$^{wt/wt}$* and *Xrcc5$^{-/-}$;Pgbd5$^{-/-}$* (**D**; t-test p = 0.0074). Bars indicate median values.

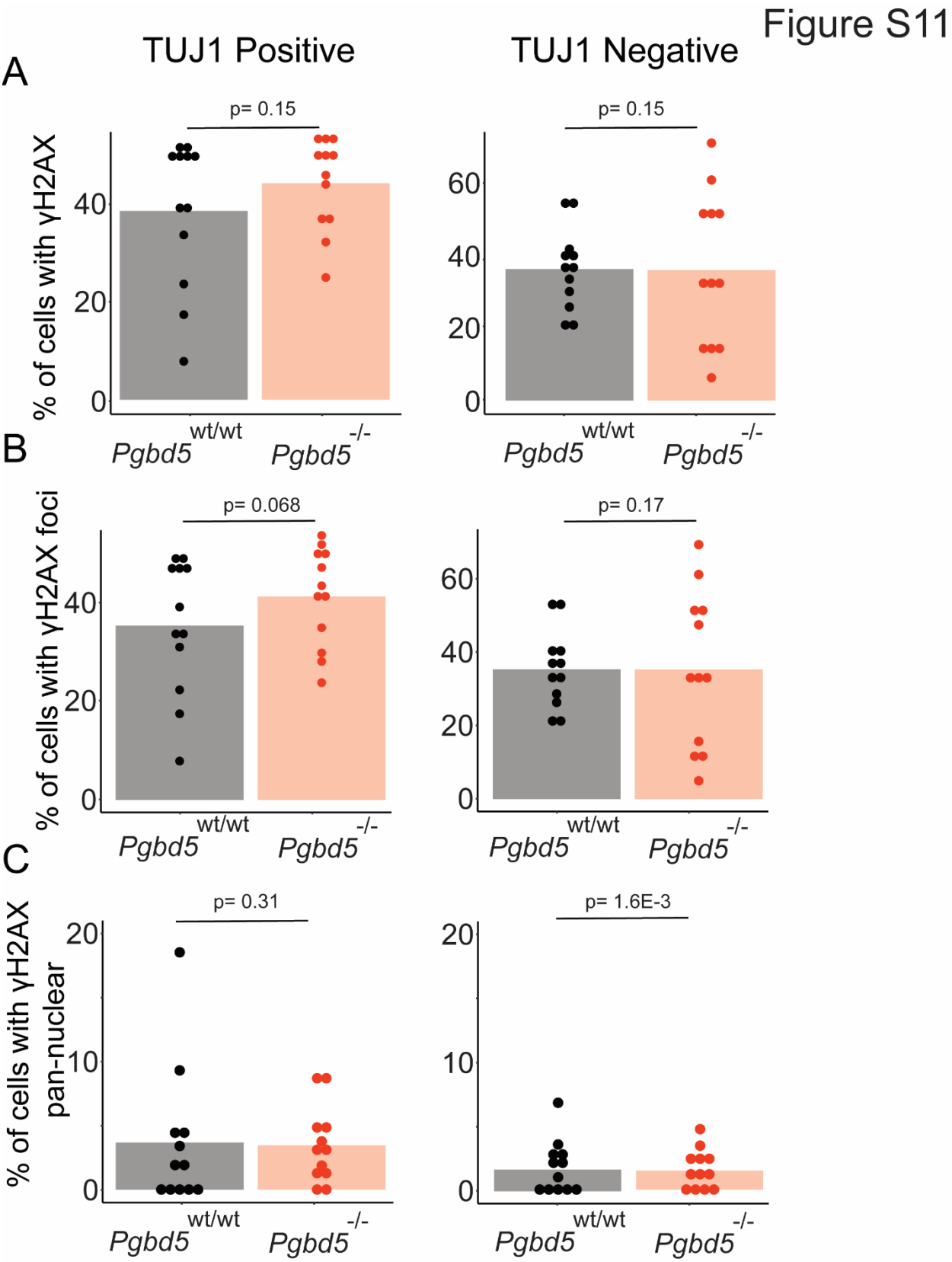

Fig. S11. *Pgbd5*-deficient E12.5 embryos lack measurable differences in brain DNA damage. **A**, Bar plots of analysis of fraction of nuclei with punctate and pan-nuclear γH2AX in postmitotic (TUJ1 positive) and proliferative (TUJ1 negative) neurons in *Pgbd5*$^{wt/wt}$ (black) and *Pgbd5*$^{-/-}$(red).

B, Bar plots of analysis of fraction of nuclei with punctate γH2AX in postmitotic (TUJ1 positive) and proliferative (TUJ1 negative) neurons in *Pgbd5*$^{wt/wt}$ (black) and *Pgbd5*$^{-/-}$(red). **C**, Bar plots of analysis of pan-nuclear γH2AX in postmitotic (TUJ1 positive) and proliferative (TUJ1 negative) neurons in *Pgbd5*$^{wt/wt}$ and *Pgbd5*$^{-/-}$. Bars indicate median values.

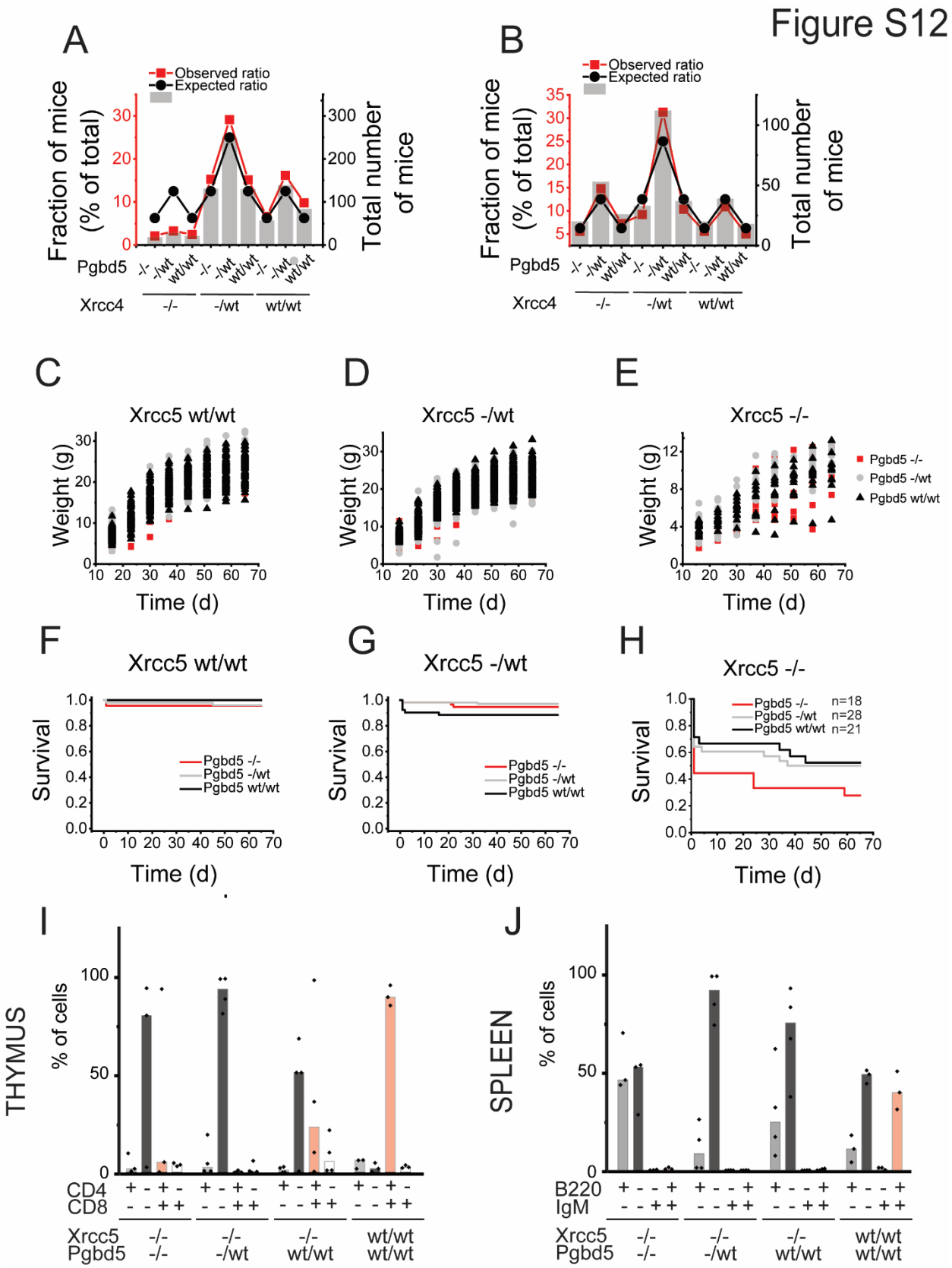

Fig. S12. *Pgbd5* deficiency does not rescue *Xrcc5*-deficiency induced immunodeficiency and exacerbates *Xrcc5*-deficiency induced runted phenotype. A-B, Mendelian ratios of *Xrcc5*$^{wt/-}$;

Pgbd5$^{wt/-}$ in-crosses of prenatal (**A**) and postnatal mice (**B**). **C-E,** Whisker plot analysis of body weight (g, grams) of *Xrcc5*$^{wt/wt}$ (**C**), *Xrcc5*$^{wt/-}$ (**D**), *Xrcc5*$^{-/-}$ (**E**) and all the different *Pgbd5* genotype mice as a function of age (d, days). **F-H,** Kaplan-Meier survival analysis of *Xrcc5*$^{wt/wt}$ (**F**), *Xrcc5*$^{wt/-}$ (**G**), *Xrcc5*$^{-/-}$ (**H**) and all the different *Pgbd5* genotype mice. **I-J**, Fluorescence activated flow cytometry analysis of immature and mature T-cells (**I**) and B-cells (**J**), isolated from thymuses and spleens of 60-day old mice. Bars represent mean values of 3 independent biological replicates.

Figure S13

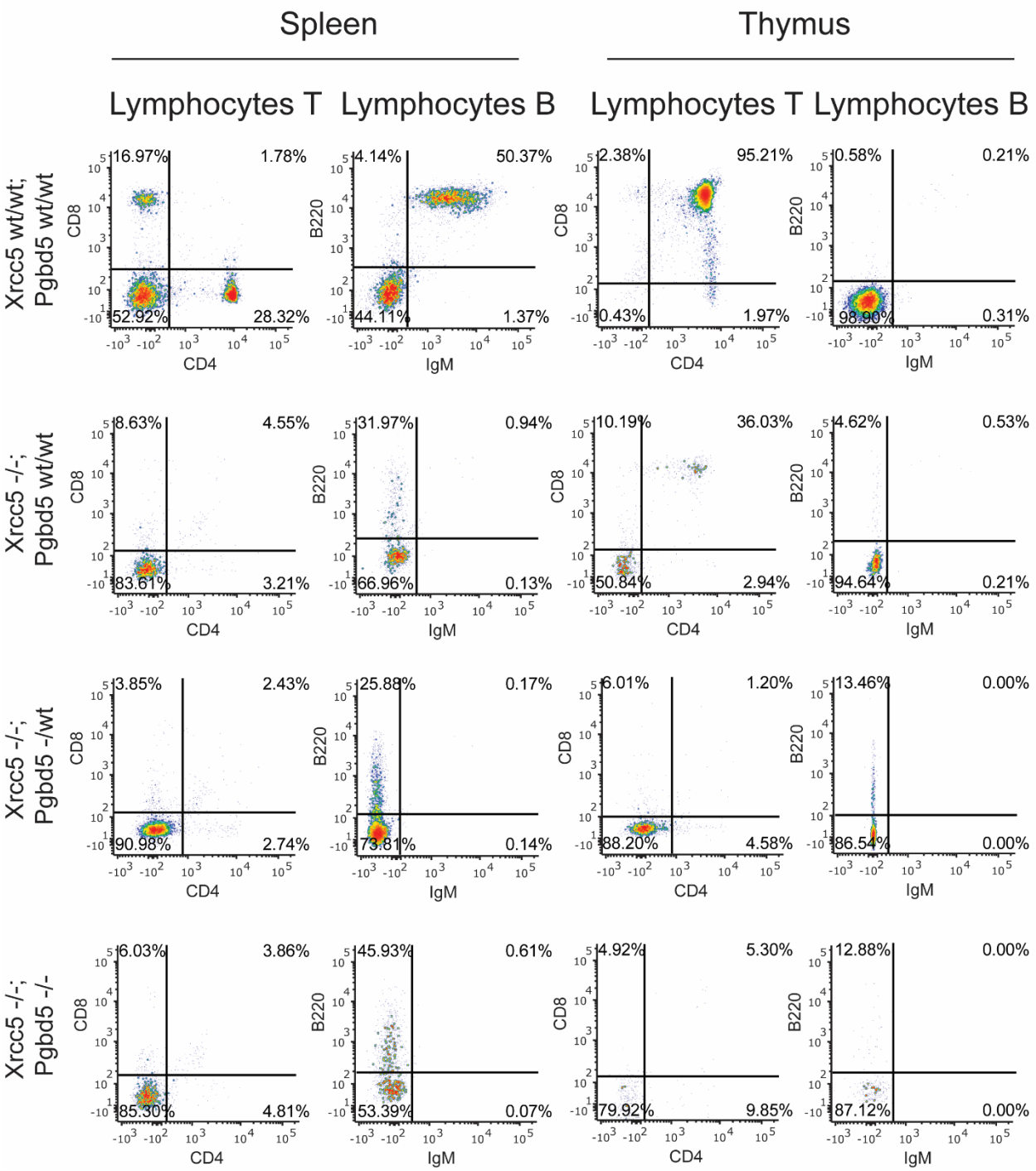

Fig. S13. *Pgbd5* deficiency does not rescue *Xrcc5*-deficiency induced immunodeficiency. Representative fluorescence activated flow cytometry plots for spleen (left) and thymus (right)

lymphocytes isolated from 60-day old mice, stained with specific T- and B-cell markers, as indicated.

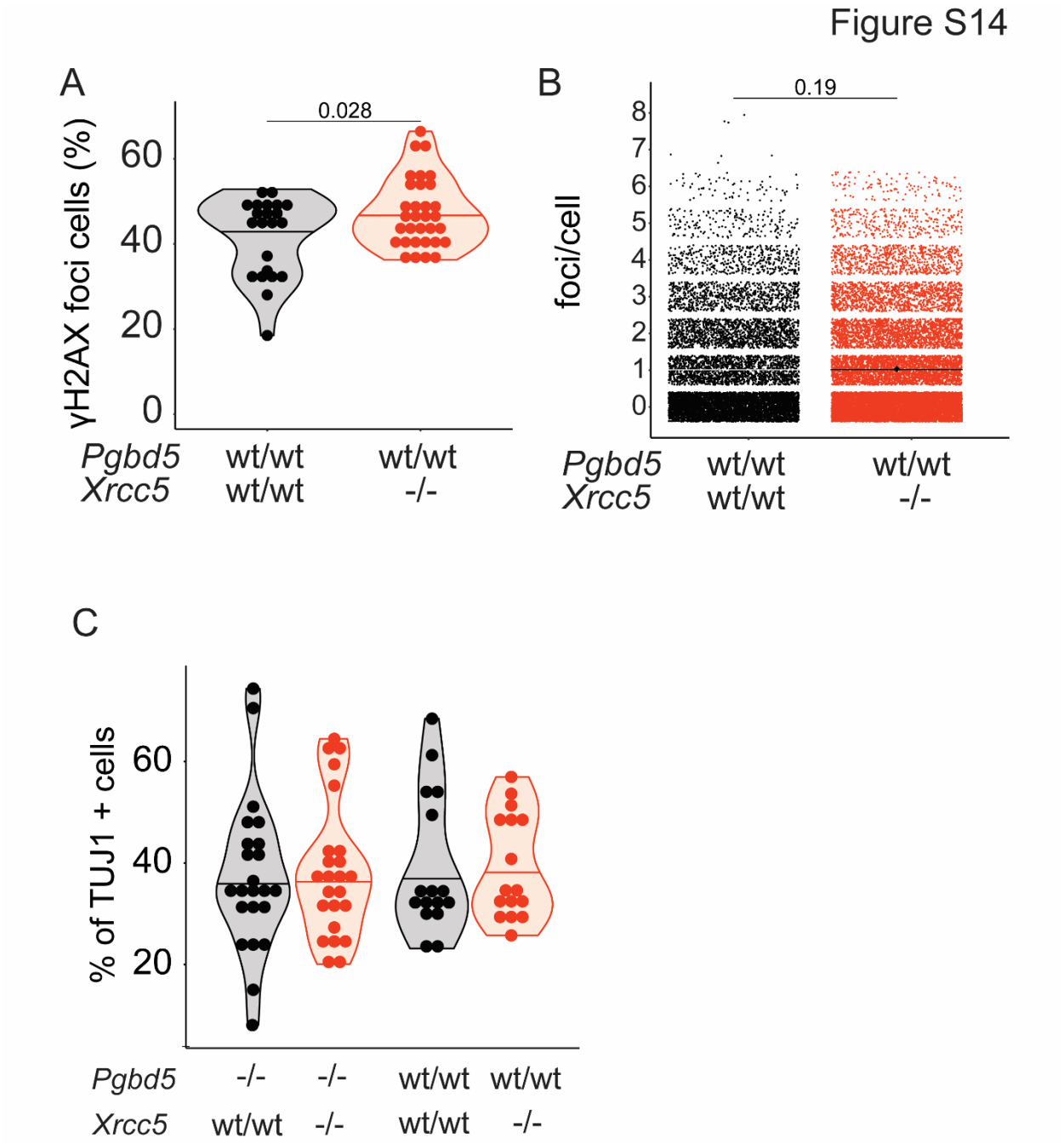

Fig. S14. *Xrcc5*$^{-/-}$ exhibit more DNA damage than their wildtype counter parts and *Pgbd5*$^{-/-}$ mice do not alter postmitotic percentage of cells. A, Violin plots of analysis of fraction of nuclei with punctate γH2AX in *Xrcc5*$^{wt/wt}$ (black) and *Xrcc5*$^{-/-}$(red) shows a significant increase on the cells containing damage (p-value=0.028). **B**,Jjitter plots showing total amount of cells with 1-8 foci/cell in *Xrcc5*$^{wt/wt}$ (black) and *Xrcc5*$^{-/-}$(red) shows no significant difference across genotypes. **C**, Violin plot for % of TUJ1 positive cells in E14.5 developing cortex shows no significant difference across *Xrcc5* and *Pgbd5* genotypes indicating that neither of them affect the postmitotic embryonic layers of the developing cortex.

Figure 16

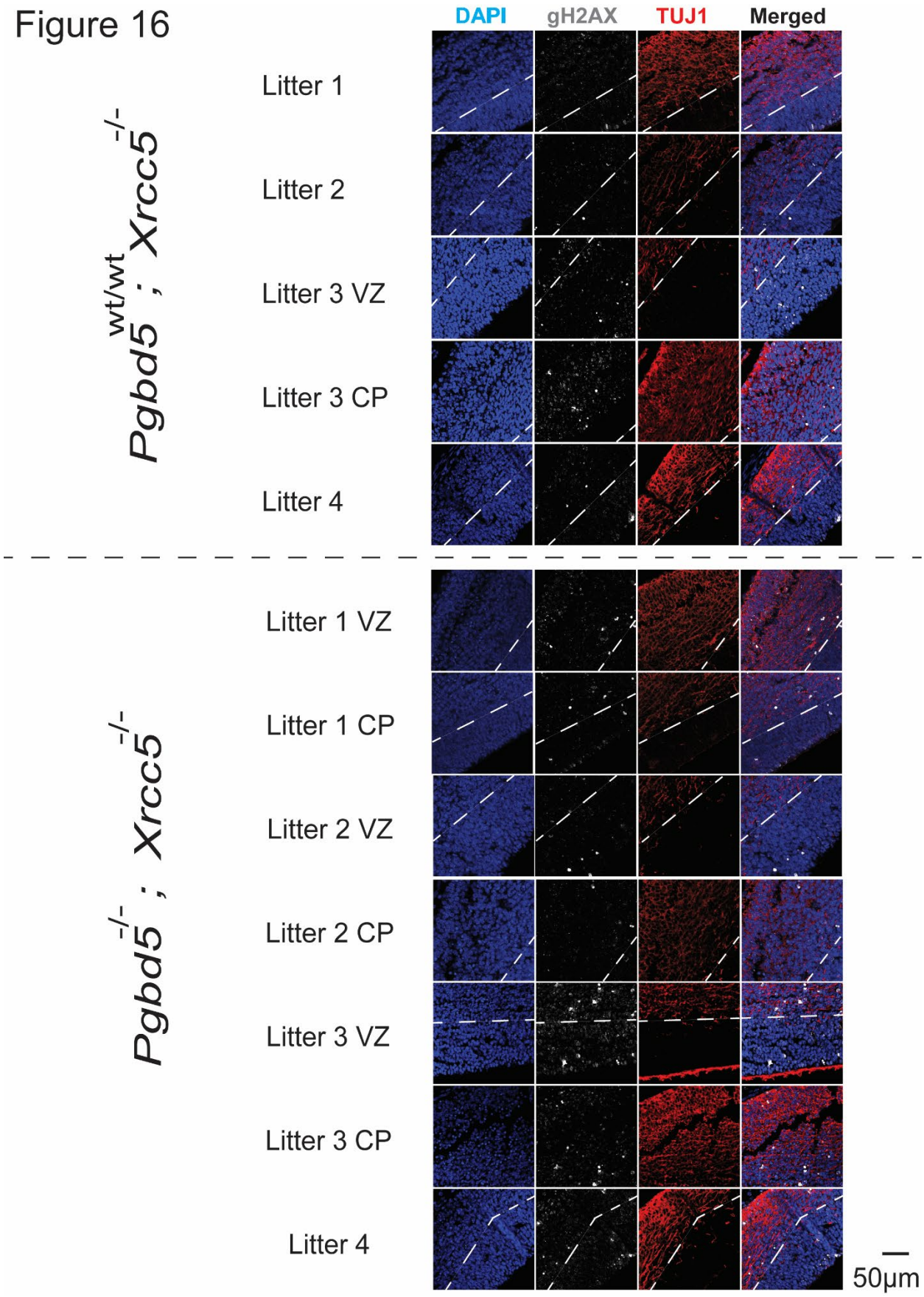
DAPI
gH2AX
TUJ1
Merged
Pgbd5wt/wt ; Xrcc5-/-
Litter 1
Litter 2
Litter 3 VZ
Litter 3 CP
Litter 4
Pgbd5-/- ; Xrcc5-/-
Litter 1 VZ
Litter 1 CP
Litter 2 VZ
Litter 2 CP
Litter 3 VZ
Litter 3 CP
Litter 4
50µm

Fig. S15. *Xrcc5*$^{-/-}$*;Pgbd5*$^{-/-}$ mice exhibit reduced DNA damage in region 1. Immunofluorescence photographs of brain region 1 from three different independent litters, stained for DNA using DAPI (blue), gH2AX (white), and Tuj1 (red), with specific litter mates indicated (VZ, ventricular zone; CP, cortical plate). Dashed lines indicate separation between ventricular zones and cortical plates. Scale bar = 50 mm.

Figure 15

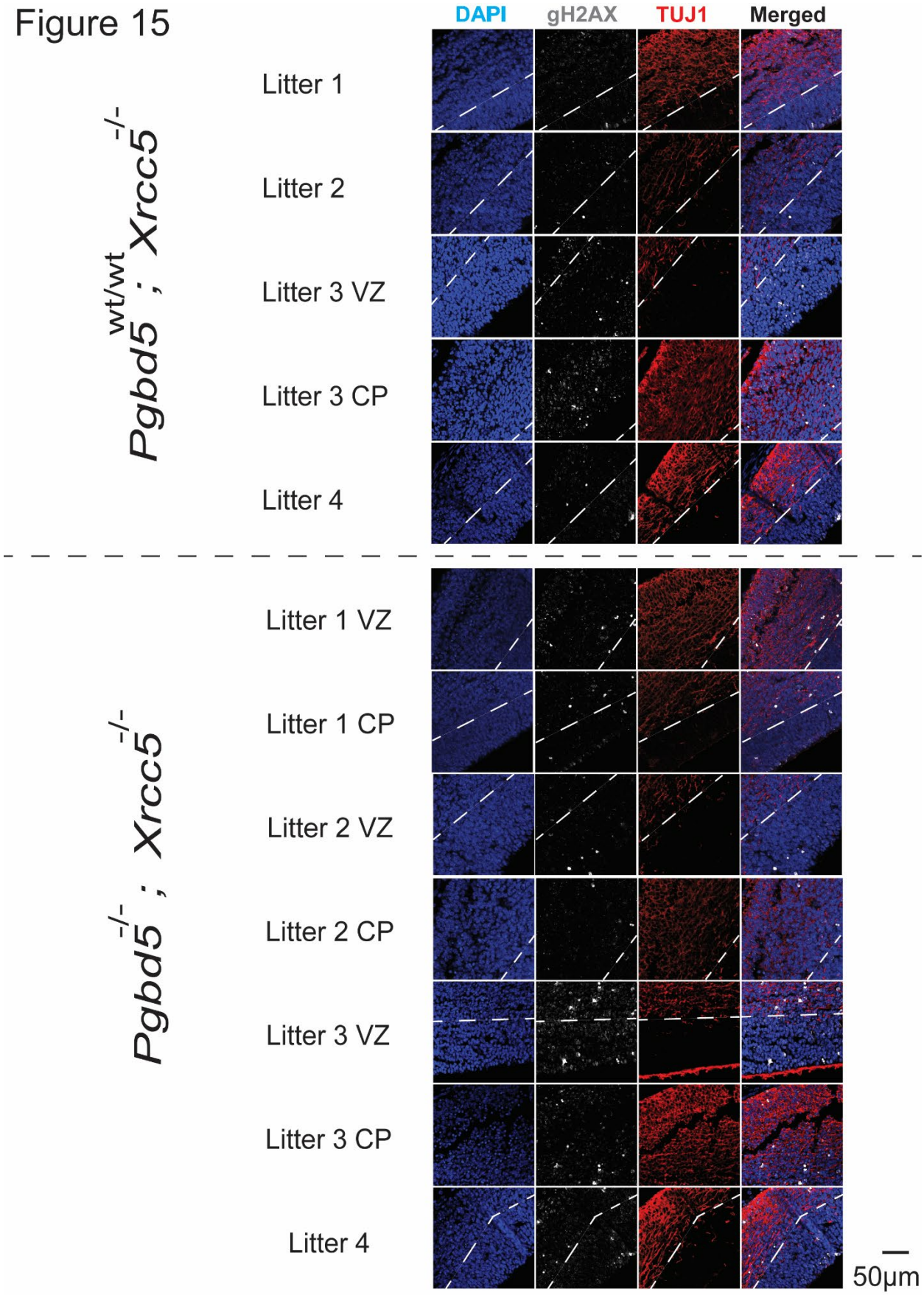
DAPI
gH2AX
TUJ1
Merged
Pgbd5wt/wt ; Xrcc5-/-
Litter 1
Litter 2
Litter 3 VZ
Litter 3 CP
Litter 4
Pgbd5-/- ; Xrcc5-/-
Litter 1 VZ
Litter 1 CP
Litter 2 VZ
Litter 2 CP
Litter 3 VZ
Litter 3 CP
Litter 4
50µm

Fig. S16. *Xrcc5*$^{-/-}$*;Pgbd5*$^{-/-}$ mice exhibit reduced DNA damage in region 2. Immunofluorescence photographs of brain region 2 from three different independent litters, stained for DNA using DAPI (blue), gH2AX (white), and Tuj1 (red), with specific litter mates indicated (VZ, ventricular zone; CP, cortical plate). Dashed lines indicate separation between ventricular zones and cortical plates. Scale bar = 50 mm.

Figure S17

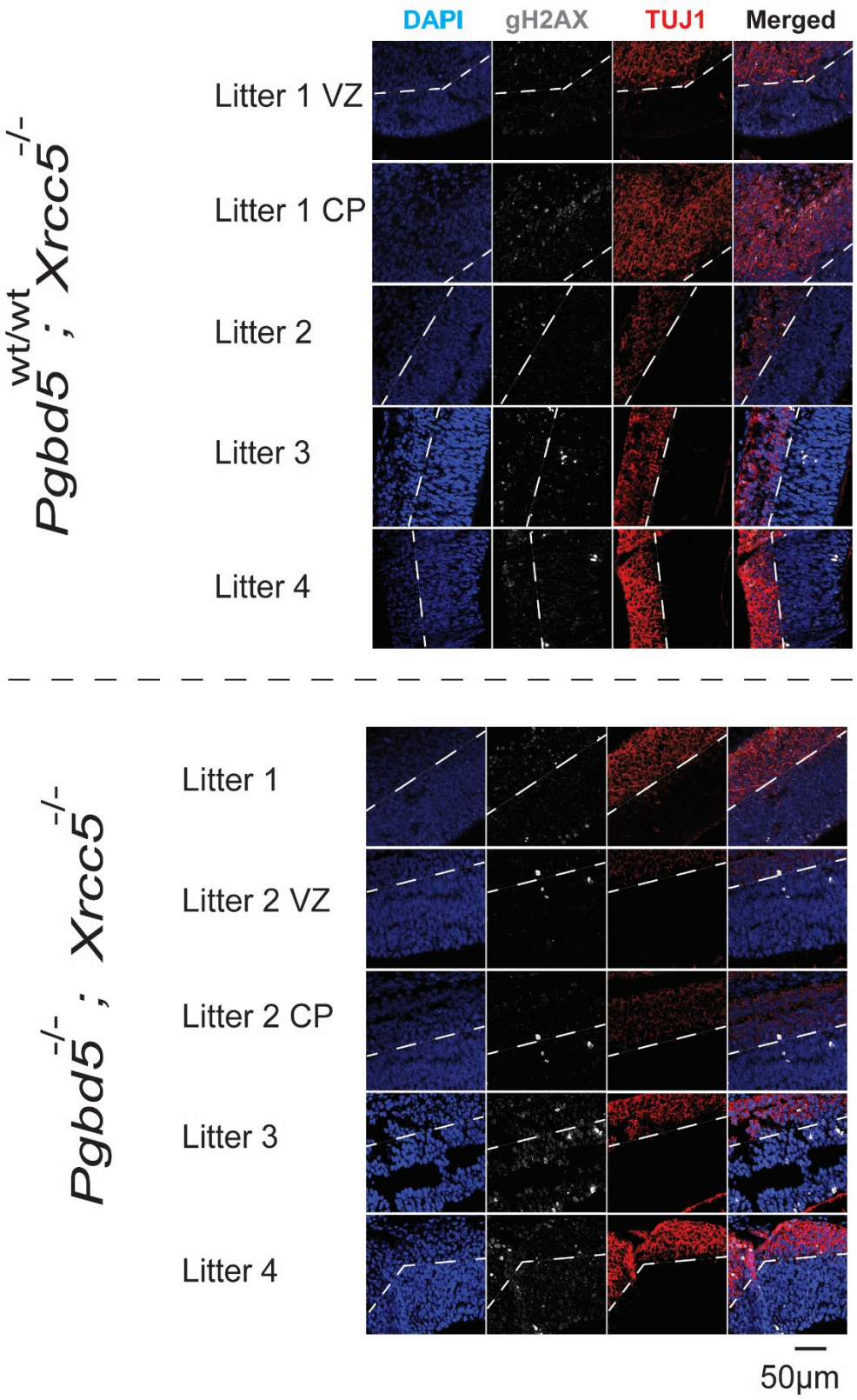
DAPI
gH2AX
TUJ1
Merged
Pgbd5 wt/wt ; Xrcc5 -/-
Litter 1 VZ
Litter 1 CP
Litter 2
Litter 3
Litter 4
Pgbd5 -/- ; Xrcc5 -/-
Litter 1
Litter 2 VZ
Litter 2 CP
Litter 3
Litter 4
50μm

Fig. S17. *Xrcc5*$^{-/-}$*;Pgbd5*$^{-/-}$ mice exhibit reduced DNA damage in region 3. Immunofluorescence photographs of brain region 3 from three different independent litters, stained for DNA using DAPI (blue), gH2AX (white), and Tuj1 (red), with specific litter mates indicated (VZ, ventricular zone; CP, cortical plate). Dashed lines indicate separation between ventricular zones and cortical plates. Scale bar = 50 mm.

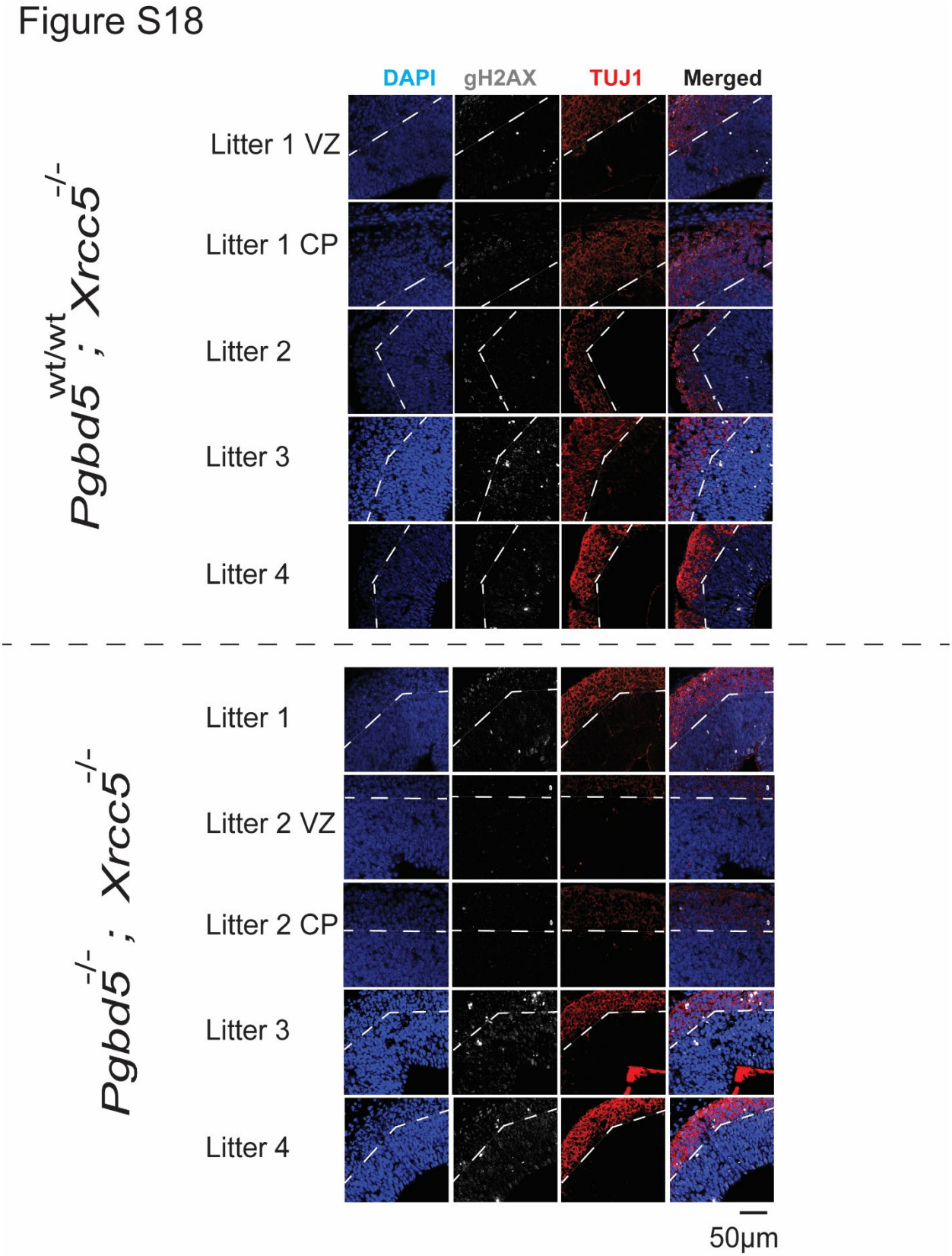

Fig. S18. *Xrcc5$^{-/-}$;Pgbd5$^{-/-}$* mice exhibit reduced DNA damage in region 4. Immunofluorescence photographs of brain region 4 from three different independent litters, stained for DNA using DAPI (blue), gH2AX (white), and TUJ1 (red), with specific litter mates indicated (VZ, ventricular zone;

CP, cortical plate). Dashed lines indicate separation between ventricular zones and cortical plates. Scale bar = 50 mm.

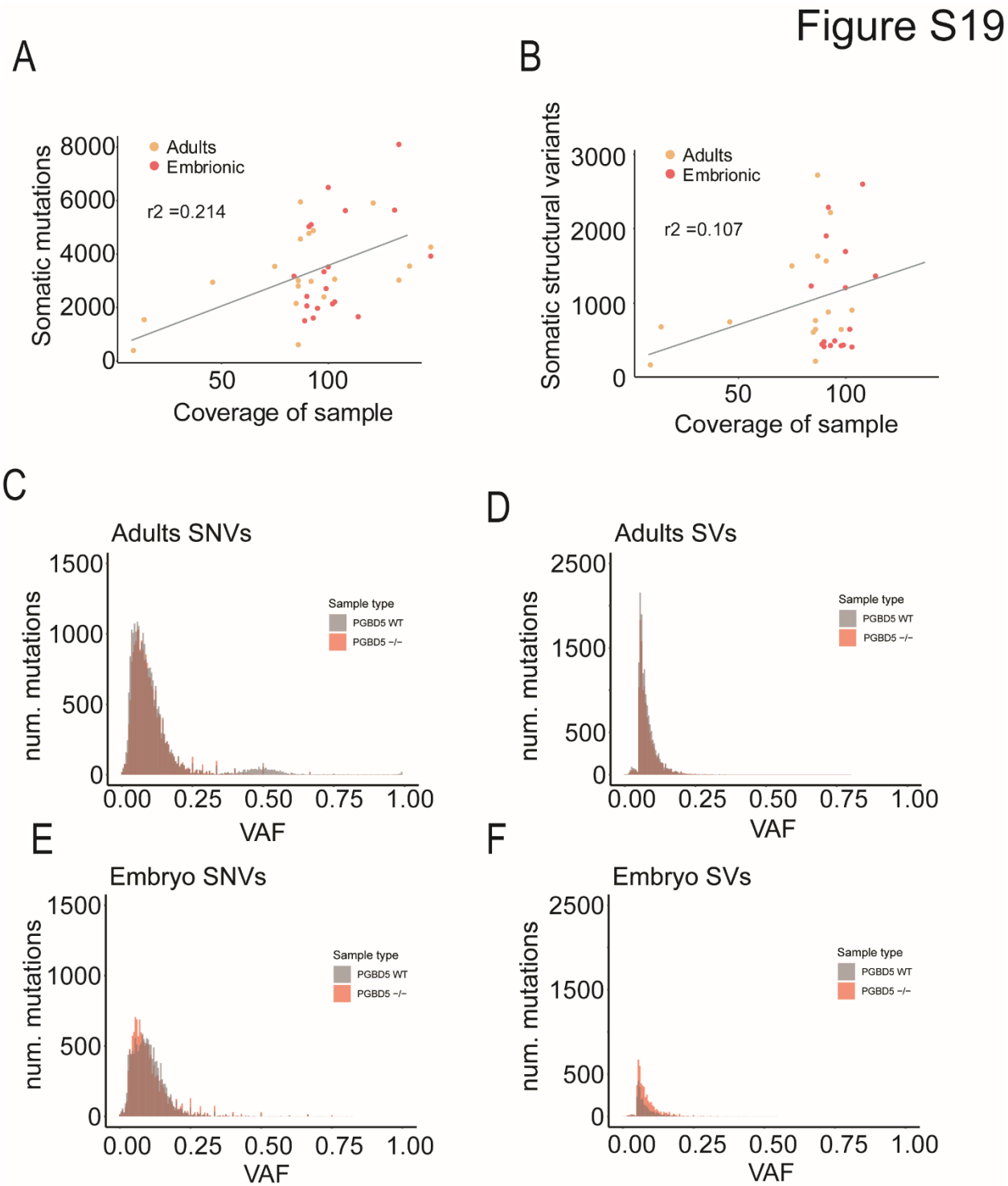

Fig. S19. WGS reliably detects infrequent variants. A-B, Sample sequencing coverage analysis shows no significant correlation between detection of somatic variants and mean sequencing coverage for somatic SNVs (**A**) and structural variants (**B**). **C-D**, Variant allele frequency (VAF) histogram for adult single nucleotide variants (SNV) (**C**) and structural variants (SV) (**D**) in *Pgbd5*$^{wt/wt}$ and *Pgbd5*$^{-/-}$. **E-F**, Variant allele frequency histogram for embryo SNV (**E**) and SV (**F**)

in *Pgbd5*$^{wt/wt}$ (grey) and *Pgbd5*$^{-/-}$(red) show that the mutations found are smaller than 0.25 indicating that they are somatic.

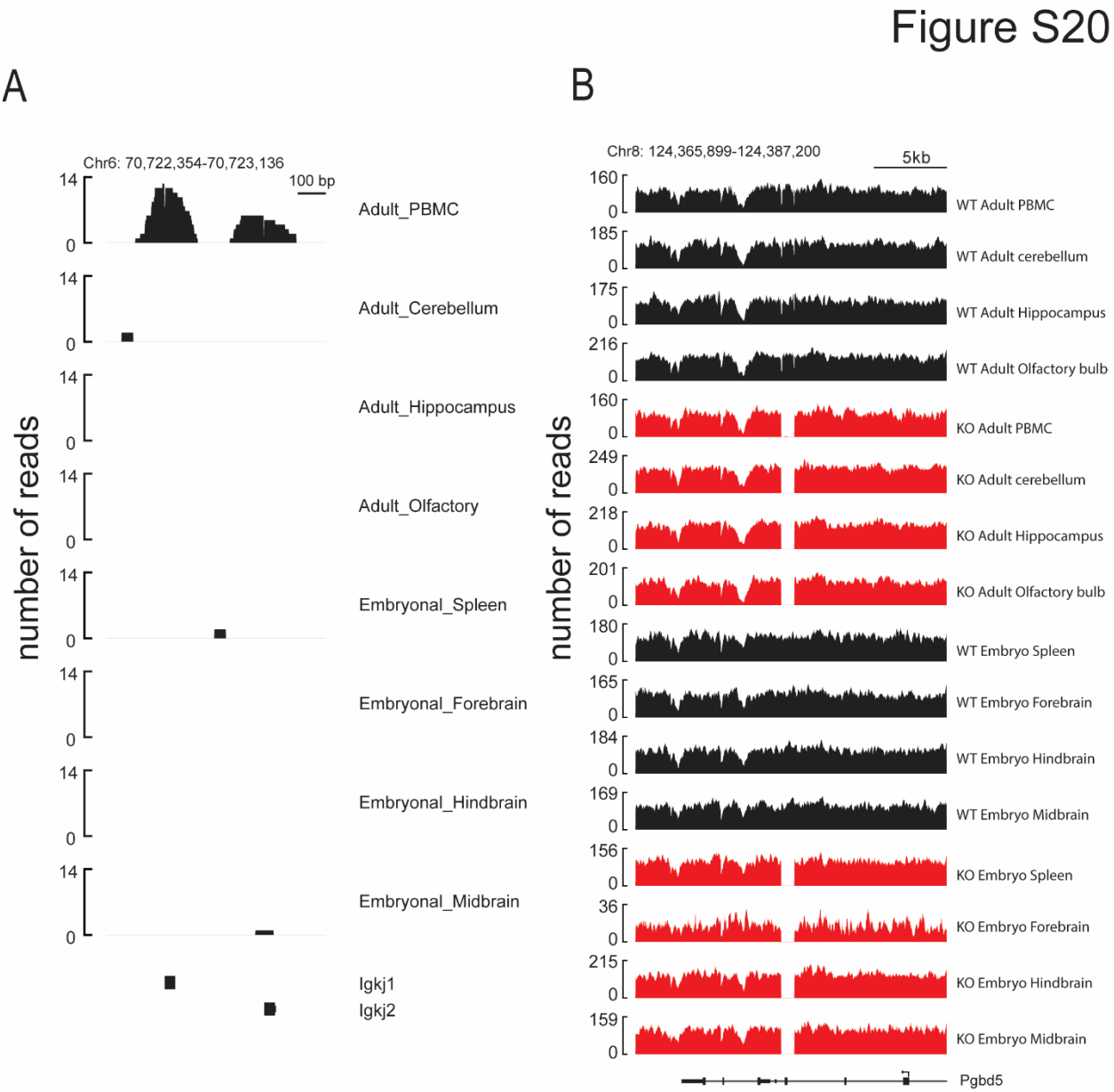

Fig. S20. Reliable detection of immunoglobulin locus rearrangements. A, Number of split reads of Igk1 and Igk2 locus in *Pgbd5*$^{wt/wt}$ adult mice show recombination only on peripheral blood mononuclear cells (PBMC) and embryo mice shows no recombination in any tissue. **B**, Split reads on Pgbd5 locus showing recombination of exon 4 across all analyzed tissues in *Pgbd5*$^{wt/wt}$ adult mice and embryo mice**.**

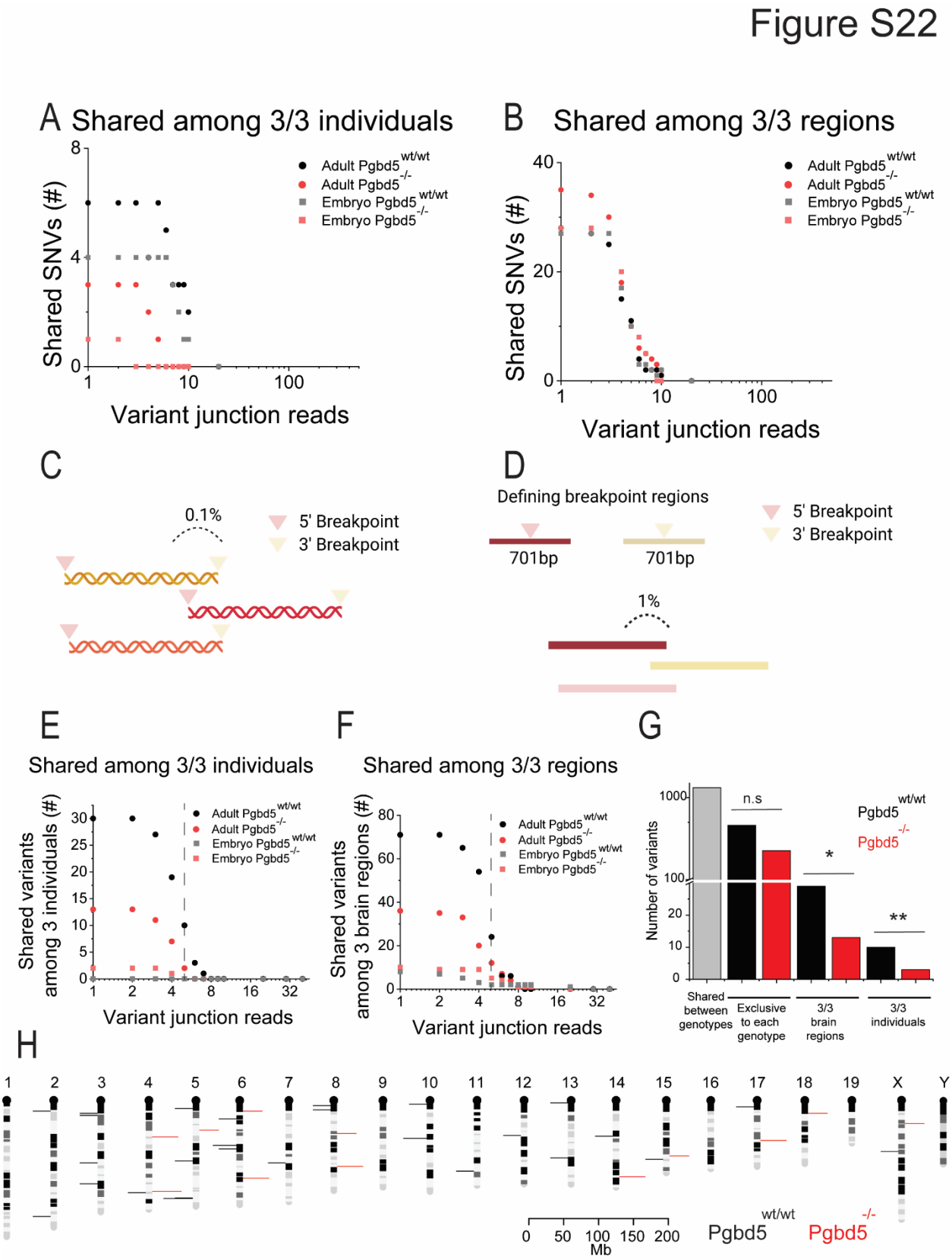

Fig. S21. Pgbd5 wildtype brains show recurrent structural variants among different individuals and brain regions. A-B, Dot plots showing number of small nucleotide variants (SNV) at different variant junction reads thresholds shared across three individuals (**A**) and three

brain regions (**B**) in adult and embryo *Pgbd5*$^{wt/wt}$ (black and grey respectively) and *Pgbd5*$^{-/-}$(red and light-red respectively). No significant difference was observed between *Pgbd5*$^{wt/wt}$ and *Pgbd5*$^{-/-}$ across the different number of supporting variant junction reads. **C**, Schematics showing the method for calculating overlap among structural variants looking at the actual genomic range (range analysis). In this case we only considered the variants that had at least an overlap of 0.1% among three individuals or brain regions. **D**, Schematics showing the method for calculating overlap among structural variants looking at the breakpoints. In this case, we defined a region of 701bp with each breakpoint on the center (breakpoint analysis). We then considered the variants that had at least an overlap of 1% among three individuals or brain regions. **E-F**, Dot plots showing numbers of recurrent somatic structural variants as a function of numbers of supporting sequencing reads shared among three individual mice (**E**) and three brain regions (**F**) in adult and embryonal brain tissues for *Pgbd5*$^{wt/wt}$ (black and grey, respectively) and *Pgbd5*$^{-/-}$ litter mate mice (red and pink, respectively). The overlap among structural variants was calculated using the range analysis: only the variants with 0.1% or more overlap were considered. Dashed line denotes allele fraction of 5 supporting reads, which minimizes most stochastic and artifactual somatic variants, as evident from their elimination in *Pgbd5*$^{-/-}$ tissues (χ^2 test p =1.6E-17 and 1.3E-9 for recurrence among different individuals and brain regions, respectively). **G**, Bar plot summarizing the results from Fig. 5**A** and **B** using the threshold of 5 reads or more. Significant differences between the number of recurrent rearrangements between *Pgbd5*$^{wt/wt}$ and *Pgbd5*$^{-/-}$ in three individuals (χ^2 p =1.6E-17) and three brain regions (χ^2 test p =1.3E-9) are observed. No significant difference is observed in non-recurrent somatic rearrangements. **H**, Mouse chromosomal ideogram showing the locations of recurrent somatic DNA rearrangements in three brain regions observed in *Pgbd5*$^{wt/wt}$ (black) and *Pgbd5*$^{-/-}$ (red) brains binned in 1 million base intervals.

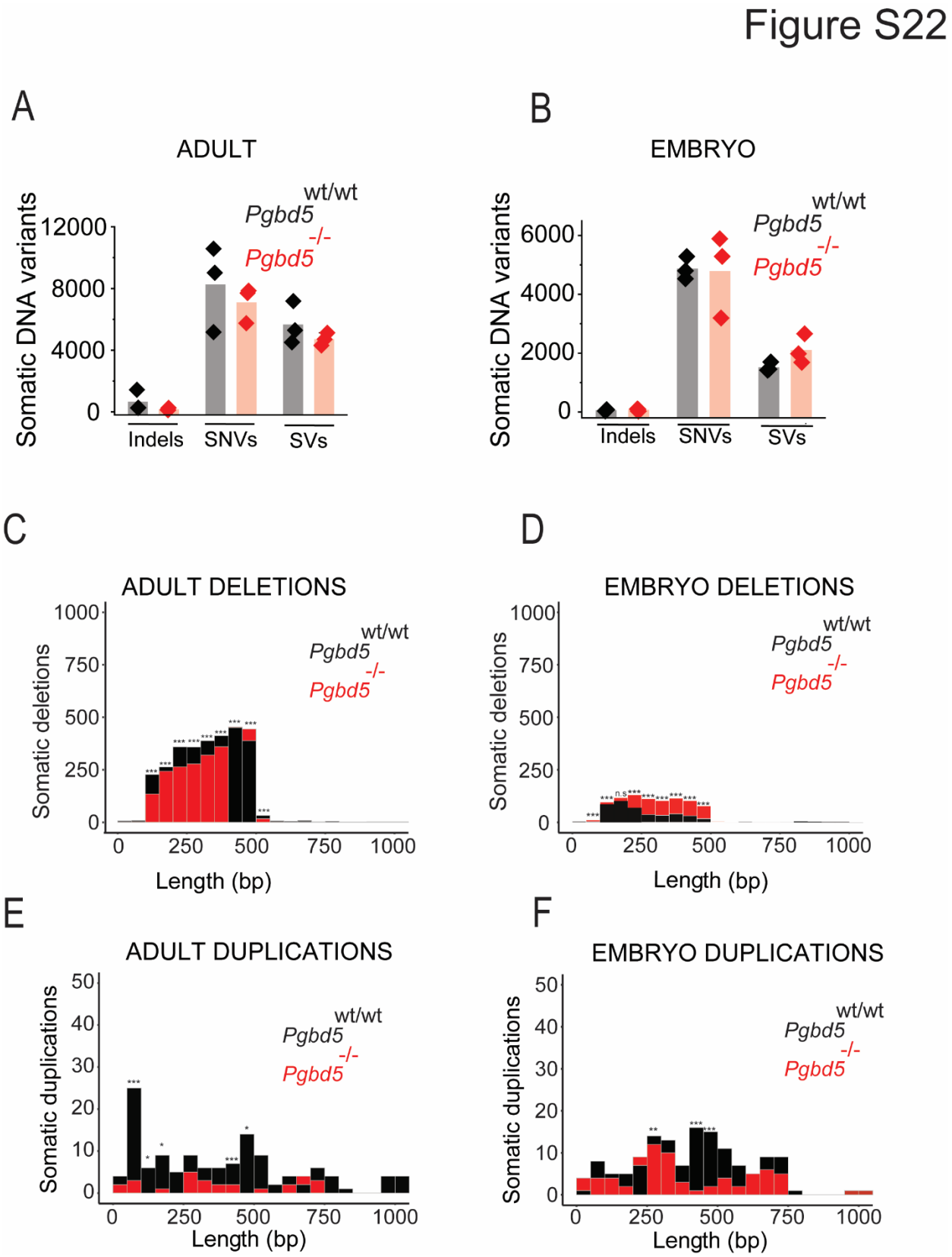

Fig. S22. Pgbd5 wildtype exclusive mutations are recurrent across different brain regions and individuals. A, Distribution of variants in *Pgbd5*$^{wt/wt}$ (black) and *Pgbd5*$^{-/-}$ (red) adult mice show no significant difference. **B**, Distribution of variants in *Pgbd5*$^{wt/wt}$ (black) and *Pgbd5*$^{-/-}$ (red)

in embryo mice show no significant difference**. C-D**, Histogram of medium-small size deletions (100-500bp) in adult and embryo *Pgbd5*$^{wt/wt}$ (grey) and *Pgbd5*$^{-/-}$(red) mice, Pearson's χ^2 test shows a significant *p*-value (1.082E-02) for 100-500bp deletions (**C**) and *Pgbd5*$^{wt/wt}$ (grey) and *Pgbd5*$^{-/-}$ (red) embryos, Pearson's χ^2 tests show a significant *p*-value (3.392E-03#) for 100-500bp duplications (**D**). **E-F,** Histogram of medium-small size duplications (100-500bp) in adult and embryo *Pgbd5*$^{wt/wt}$ (grey) and *Pgbd5*$^{-/-}$(red) mice, Pearson's χ^2 test shows a tendency *p*-value (0.35) for 250-500bp duplications in adult mice (**E**) and *Pgbd5*$^{wt/wt}$ (grey) and *Pgbd5*$^{-/-}$(red) embryos, Pearson's χ^2 test shows a trend *p*-value (0.12) for 150-500bp duplications in embryos (**F**). SNVs= small nucleotide variants, SVs= structural variants.

Supplemental Data Files

Data S1 (separate file)

Pgbd5_KI_sequencing: contains a table listing the structural variants found in whole genome sequencing of Pgbd5 KI mice founders

Data S2 (separate file)

Quantified_images: contain all the original files that were used to manually quantify the DNA damage of *Pgbd5*$^{wt/wt}$, *Pgbd5*$^{-/-}$, *Xrcc5*$^{-/-}$;*Pgbd5*$^{wt/wt}$ and *Xrcc5*$^{-/-}$; *Pgbd5*$^{-/-}$ E14.5 embryos.

Data S3 (separate file)

WGS_data: contains files and information related to the whole genome sequencing of *Pgbd5*$^{wt/wt}$ and *Pgbd5*$^{-/-}$ mice, including metadata of SRA deposit of the whole genome sequencing data, the generated vcf files used in the analysis as well as a table listing the recurrent rearrangements found in *Pgbd5*$^{wt/wt}$.

Supplemental References

1. G. M. Edelman, Neural Darwinism: selection and reentrant signaling in higher brain function. *Neuron* **10**, 115-125 (1993).
2. D. E. Barnes, G. Stamp, I. Rosewell, A. Denzel, T. Lindahl, Targeted disruption of the gene encoding DNA ligase IV leads to lethality in embryonic mice. *Current biology : CB* **8**, 1395-1398 (1998).
3. Y. Gao *et al.*, A targeted DNA-PKcs-null mutation reveals DNA-PK-independent functions for KU in V(D)J recombination. *Immunity* **9**, 367-376 (1998).
4. K. M. Frank *et al.*, Late embryonic lethality and impaired V(D)J recombination in mice lacking DNA ligase IV. *Nature* **396**, 173-177 (1998).
5. J. M. Sekiguchi *et al.*, Nonhomologous end-joining proteins are required for V(D)J recombination, normal growth, and neurogenesis. *Cold Spring Harb Symp Quant Biol* **64**, 169-181 (1999).
6. Y. Lee, D. E. Barnes, T. Lindahl, P. J. McKinnon, Defective neurogenesis resulting from DNA ligase IV deficiency requires Atm. *Genes Dev* **14**, 2576-2580 (2000).
7. Y. Gu *et al.*, Defective embryonic neurogenesis in Ku-deficient but not DNA-dependent protein kinase catalytic subunit-deficient mice. *Proc Natl Acad Sci U S A* **97**, 2668-2673 (2000).
8. K. M. Frank *et al.*, DNA ligase IV deficiency in mice leads to defective neurogenesis and embryonic lethality via the p53 pathway. *Mol Cell* **5**, 993-1002 (2000).
9. P. J. McKinnon, Maintaining genome stability in the nervous system. *Nat Neurosci* **16**, 1523-1529 (2013).
10. W. J. Dreyer, W. R. Gray, L. Hood, The Genetic, Molecular, and Cellular Basis of Antibody Formation: Some Facts and a Unifying Hypothesis. *Cold Spring Harbor Symposia on Quantitative Biology* **32**, 353-367 (1967).
11. Q. Wu, T. Maniatis, A striking organization of a large family of human neural cadherin-like cell adhesion genes. *Cell* **97**, 779-790 (1999).
12. X. Wang *et al.*, Gamma protocadherins are required for survival of spinal interneurons. *Neuron* **36**, 843-854 (2002).
13. P. C. Wei *et al.*, Long Neural Genes Harbor Recurrent DNA Break Clusters in Neural Stem/Progenitor Cells. *Cell* **164**, 644-655 (2016).
14. F. W. Alt, B. Schwer, DNA double-strand breaks as drivers of neural genomic change, function, and disease. *DNA Repair (Amst)* **71**, 158-163 (2018).
15. A. M. D'Gama, C. A. Walsh, Somatic mosaicism and neurodevelopmental disease. *Nat Neurosci* **21**, 1504-1514 (2018).
16. I. L. Weissman, F. H. Gage, A Mechanism for Somatic Brain Mosaicism. *Cell* **164**, 593-595 (2016).
17. C. Sun *et al.*, Mapping the Complex Genetic Landscape of Human Neurons. *bioRxiv*, (2023).
18. J. Kim *et al.*, Prevalence and mechanisms of somatic deletions in single human neurons during normal aging and in DNA repair disorders. *Nature communications* **13**, 5918 (2022).
19. A. G. Henssen *et al.*, Genomic DNA transposition induced by human PGBD5. *eLife* **4**, e10565 (2015).
20. A. G. Henssen *et al.*, PGBD5 promotes site-specific oncogenic mutations in human tumors. *Nat Genet* **49**, 1005-1014 (2017).

21. R. Keller *et al.*, Childhood cancer mutagenesis caused by a domesticated DNA transposase. *bioRxiv*, 2022.2007.2005.498128 (2022).
22. L. Helou *et al.*, The piggyBac-derived protein 5 (PGBD5) transposes both the closely and the distantly related piggyBac-like elements Tcr-pble and Ifp2. *Journal of molecular biology* **433**, 166839 (2021).
23. Y. Bigot *et al.*, Analysis of DNA transposition by DNA transposases in human cells. *bioRxiv*, 2023.2004.2026.538406 (2023).
24. A. G. Henssen *et al.*, Therapeutic targeting of PGBD5-induced DNA repair dependency in pediatric solid tumors. *Science translational medicine* **9**, (2017).
25. N. Sobreira, F. Schiettecatte, D. Valle, A. Hamosh, GeneMatcher: a matching tool for connecting investigators with an interest in the same gene. *Human mutation* **36**, 928-930 (2015).
26. M. Osterwalder *et al.*, Dual RMCE for efficient re-engineering of mouse mutant alleles. *Nat Methods* **7**, 893-895 (2010).
27. M. Götz, W. B. Huttner, The cell biology of neurogenesis. *Nature Reviews Molecular Cell Biology* **6**, 777-788 (2005).
28. J. R. Menezes, M. B. Luskin, Expression of neuron-specific tubulin defines a novel population in the proliferative layers of the developing telencephalon. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **14**, 5399-5416 (1994).
29. L. J. Mah, A. El-Osta, T. C. Karagiannis, gammaH2AX: a sensitive molecular marker of DNA damage and repair. *Leukemia* **24**, 679-686 (2010).
30. T. A. Rinehart, C. Dean, C. F. Weil, Comparative analysis of non-random DNA repair following Ac transposon excision in maize and Arabidopsis. *The Plant journal : for cell and molecular biology* **12**, 1419-1427 (1997).
31. L. E. Woodard *et al.*, Comparative analysis of the recently discovered hAT transposon TcBuster in human cells. *PLoS One* **7**, e42666 (2012).
32. A. Nussenzweig *et al.*, Requirement for Ku80 in growth and immunoglobulin V(D)J recombination. *Nature* **382**, 551-555 (1996).
33. C. Gawad, W. Koh, S. R. Quake, Single-cell genome sequencing: current state of the science. *Nat Rev Genet* **17**, 175-188 (2016).
34. T. Rausch *et al.*, DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333-i339 (2012).
35. A. McKenna *et al.*, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303 (2010).
36. K. Ye, M. H. Schulz, Q. Long, R. Apweiler, Z. Ning, Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865-2871 (2009).
37. H. Igarashi, S. C. Gregory, T. Yokota, N. Sakaguchi, P. W. Kincade, Transcription from the RAG1 locus marks the earliest lymphocyte progenitors in bone marrow. *Immunity* **17**, 117-130 (2002).
38. A. V. Nesta, D. Tafur, C. R. Beck, Hotspots of Human Mutation. *Trends Genet* **37**, 717-729 (2021).
39. W. R. Gray, W. J. Dreyer, L. Hood, Mechanism of antibody synthesis: size differences between mouse kappa chains. *Science* **155**, 465-467 (1967).
40. A. J. Blaschke, J. A. Weiner, J. Chun, Programmed cell death is a universal feature of embryonic and postnatal neuroproliferative regions throughout the central nervous system. *J Comp Neurol* **396**, 39-50 (1998).

41. S. Rakic, N. Zecevic, Programmed cell death in the developing human telencephalon. *Eur J Neurosci* **12**, 2721-2734 (2000).
42. L. Lossi, A. Merighi, In vivo cellular and molecular mechanisms of neuronal apoptosis in the mammalian CNS. *Prog Neurobiol* **69**, 287-312 (2003).
43. W. Wu *et al.*, Neuronal enhancers are hotspots for DNA single-strand break repair. *Nature* **593**, 440-444 (2021).
44. G. Pascarella *et al.*, Recombination of repeat elements generates somatic complexity in human genomes. *Cell* **185**, 3025-3040.e3026 (2022).
45. V. Billon *et al.*, Somatic retrotransposition in the developing rhesus macaque brain. *Genome Res* **32**, 1298-1314 (2022).
46. D. G. Schatz, Developing B-cell theories. *Nature* **400**, 614-615, 617 (1999).
47. S. Keeney, J. Lange, N. Mohibullah, Self-organization of meiotic recombination initiation: general principles and molecular pathways. *Annu Rev Genet* **48**, 187-214 (2014).
48. N. Rajaei, K. K. Chiruvella, F. Lin, S. U. Aström, Domesticated transposase Kat1 and its fossil imprints induce sexual differentiation in yeast. *Proc Natl Acad Sci U S A* **111**, 15491-15496 (2014).
49. J. Bischerour *et al.*, Six domesticated PiggyBac transposases together carry out programmed DNA elimination in Paramecium. *eLife* **7**, (2018).
50. A. Kentsis, Why do young people get cancer? *Pediatr Blood Cancer* **67**, e28335 (2020).
51. C. Garel *et al.*, Biometry of the corpus callosum in children: MR imaging reference data. *AJNR. American journal of neuroradiology* **32**, 1436-1443 (2011).
52. A. M. Pagnozzi *et al.*, Identifying relevant biomarkers of brain injury from structural MRI: Validation using automated approaches in children with unilateral cerebral palsy. *PLoS One* **12**, e0181605 (2017).
53. T. N. Ferraro *et al.*, Mapping Loci for Pentylenetetrazol-Induced Seizure Susceptibility in Mice. *The Journal of Neuroscience* **19**, 6733-6739 (1999).
54. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
55. T. Rausch, M. Hsi-Yang Fritz, J. O. Korbel, V. Benes, Alfred: interactive multi-sample BAM alignment statistics, feature counting and feature annotation for long- and short-read sequencing. *Bioinformatics* **35**, 2489-2491 (2018).
56. K. M. Raine *et al.*, cgpPindel: Identifying Somatically Acquired Insertion and Deletion Events from Paired End Sequencing. *Curr Protoc Bioinformatics* **52**, 15.17.11-15.17.12 (2015).
57. K. Ye, M. H. Schulz, Q. Long, R. Apweiler, Z. Ning, Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865-2871 (2009).
58. A. R. Quinlan, BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr Protoc Bioinformatics* **47**, 11.12.11-34 (2014).
59. W. McLaren *et al.*, The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122 (2016).
60. G. G. Faust, I. M. Hall, SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics* **30**, 2503-2505 (2014).
61. S. Kurtenbach, J. William Harbour, SparK: A Publication-quality NGS Visualization Tool. *bioRxiv*, 845529 (2019).